Statistical Analysis of Social Networks

Krista J. Gile
University of Massachusetts, Amherst

Octover 24, 2013

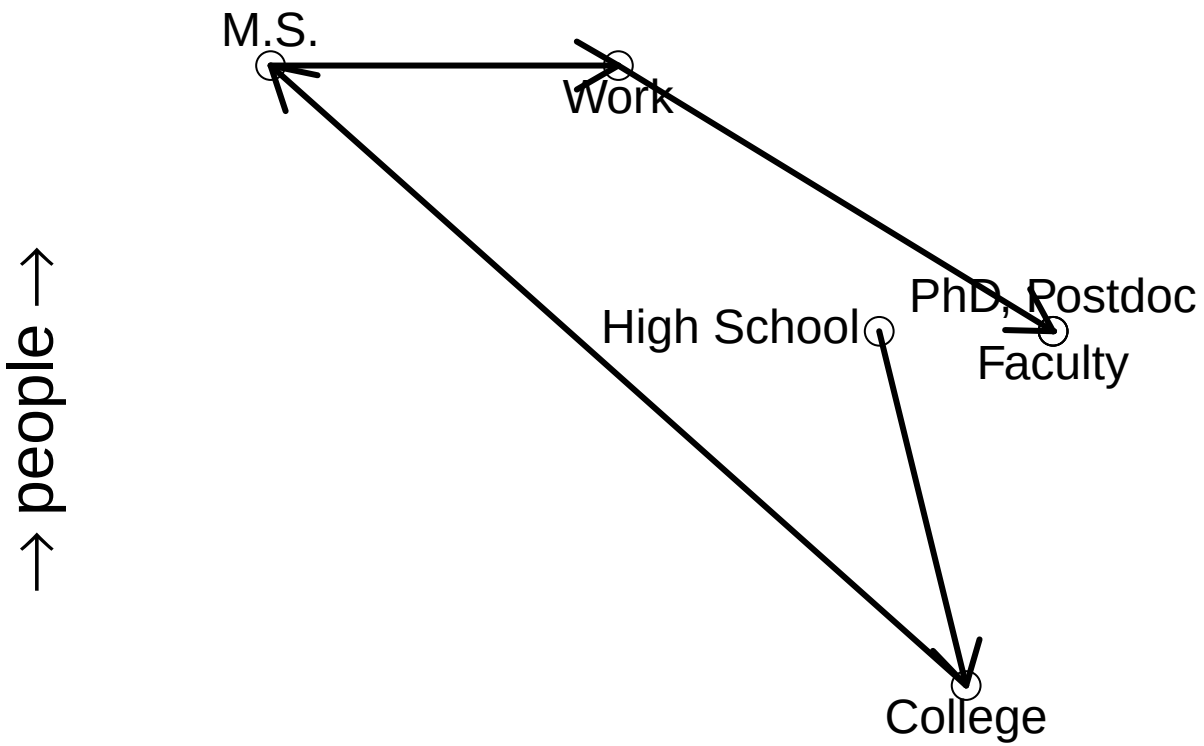
Collaborators:

- Isabelle Beaudry, UMass Amherst
- Elena Erosheva, University of Washington
- Ian Fellows, FellStat
- Karen Fredriksen-Goldsen, University of Washington
- Krista J. Gile, UMass, Amherst
- Mark S. Handcock, UCLA
- Lisa G. Johnston, Tulane University, UCSF
- Corinne M. Mar, University of Washington
- Miles Ott, Carleton College
- Matt Salganik, Princeton University
- Amber Tomas, Mathematica Policy Research

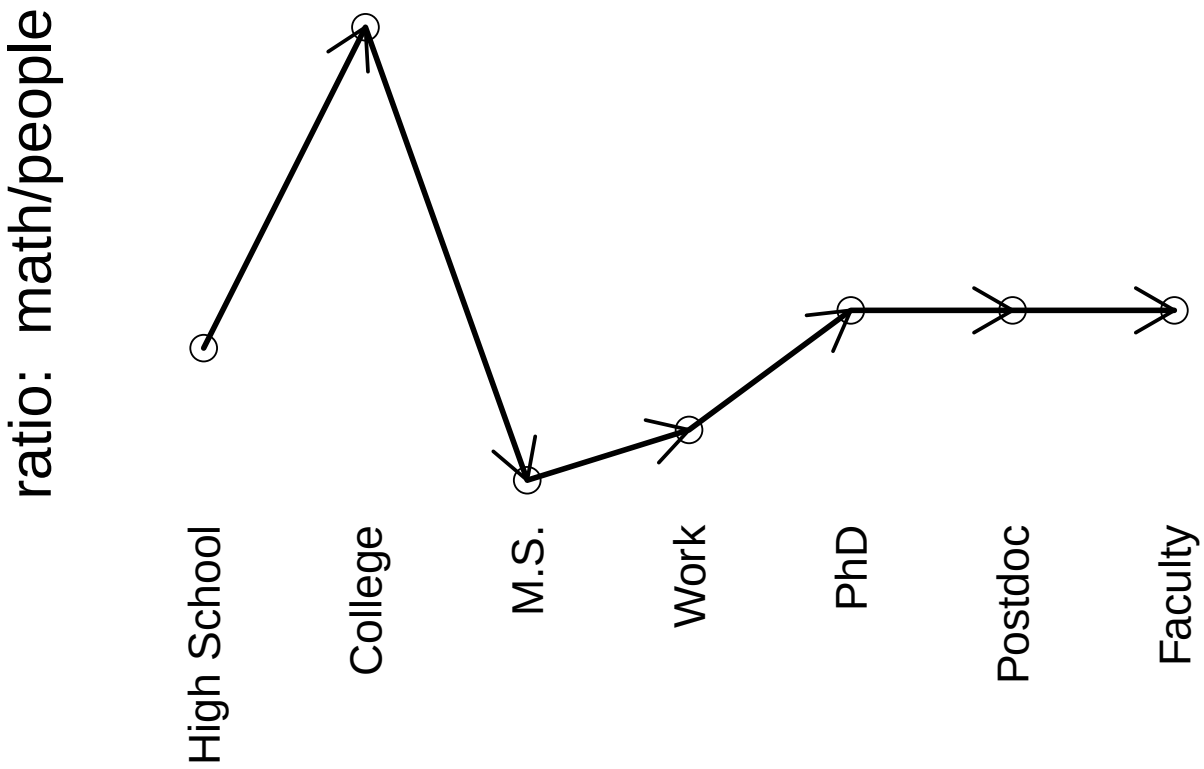
For details, see:

- http://www.math.umass.edu/˜ gile

Career Path (how it felt)



Career Path (employer version)



Outline of Presentation

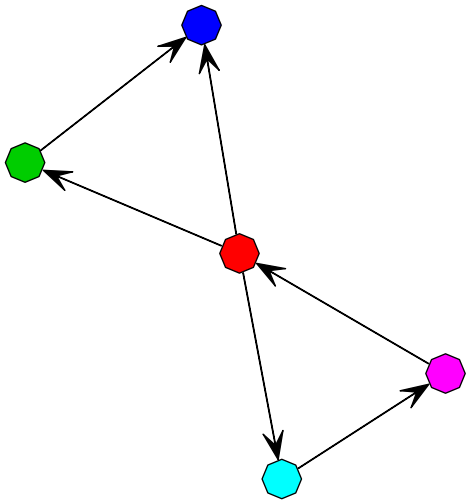
1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

(Cross-Sectional) Social Networks

- Social Network: Tool to formally represent and quantify relational social structure.
- Relations can include: friendships, workplace collaborations, international trade
- Represent mathematically as a sociomatrix, Y, where
  $Y_{ij} =$ the value of the relationship from i to j



(c) Sociogram

	0	1	1	1	0
	0	0	1	0	0
	0	0	0	0	0
	0	0	0	0	1
	1	0	0	0	0

(d) Sociomatrix

Social Networks that are harder:

- Networks that change over time (ties, nodes, discrete, continuous, models, data)
- (effectively) Unbounded networks (the internet, social contacts in the world)
- Multiple social networks (friendships, co-work, leisure)
- Multi-modal networks (people and places)
- Valued networks (amount of trade, signed networks)
- Event networks (e-mail, armed conflict)

Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

Why do we want to analyze social networks?

- Because it's cool and fun.
- Because we want information we can't get other ways.
  - Learn about nodes
  - Learn about relations
  - Learn about processes

- Learn about the particular network
- Learn about the process the network came from

Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

What can we know about a social network?

- All nodes, all relations
- All nodes, some relations
- Some nodes, some relations

Others we won't emphasize here:

- Attributes of nodes (where do your facebook contacts live?)
- Attributes of relations (who is your *best* friend?)
- Information on processes over the network (how much trade between countries?)

Partially-Observed Social Network Data

Some portion of the social network is often unobserved.

$Y =$

	A	B	C	D
A	-	1	0	0
B	0	-	1	1
C	0	0	-	0
D	1	1	1	-

$Y_{obs} =$

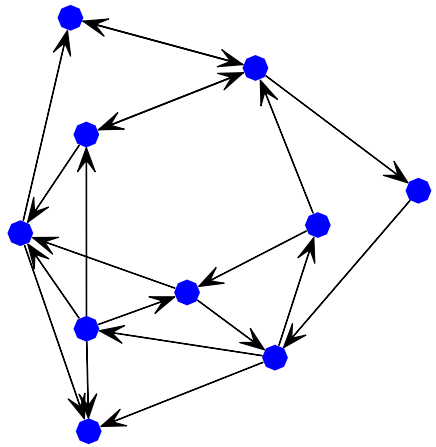
	A	B	C	D
A	-	?	?	?
B	?	-	?	?
C	0	0	-	0
D	1	1	1	-

$D =$

	A	B	C	D
A	-	0	0	0
B	0	-	0	0
C	1	1	-	1
D	1	1	1	-

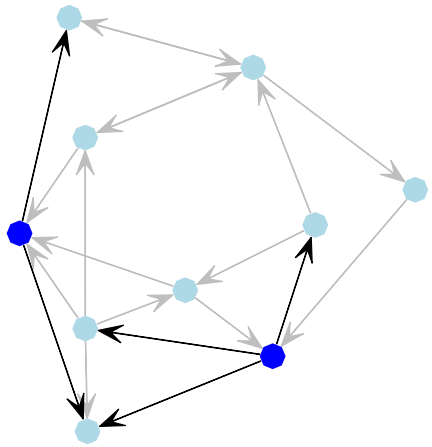
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



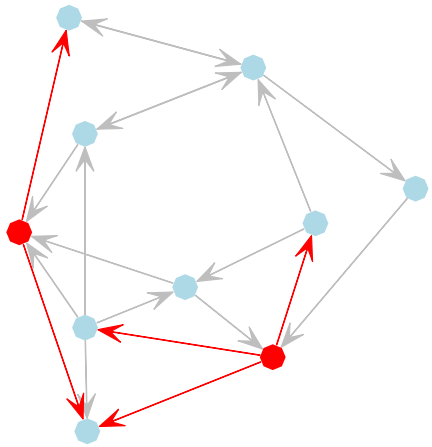
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



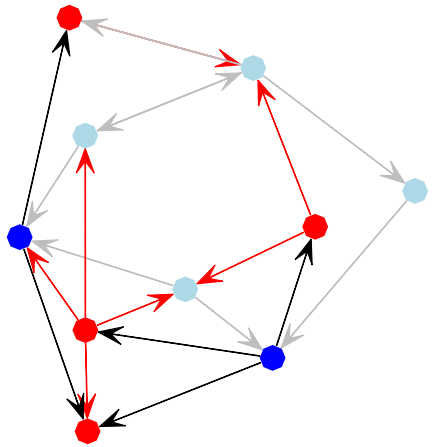
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



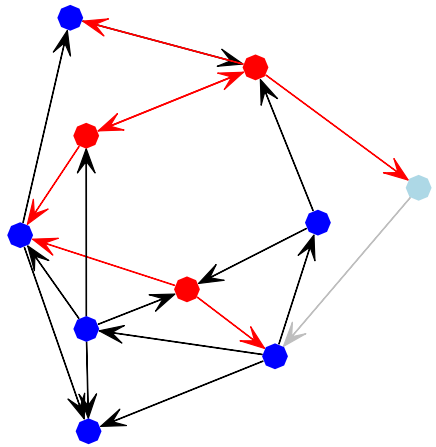
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



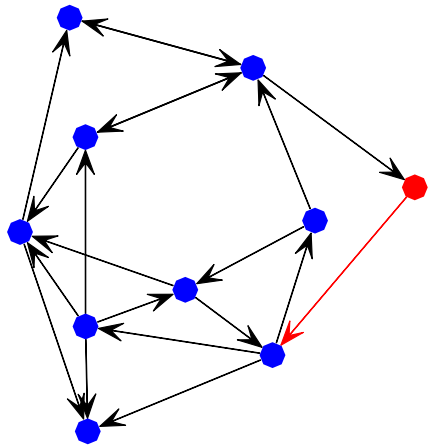
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



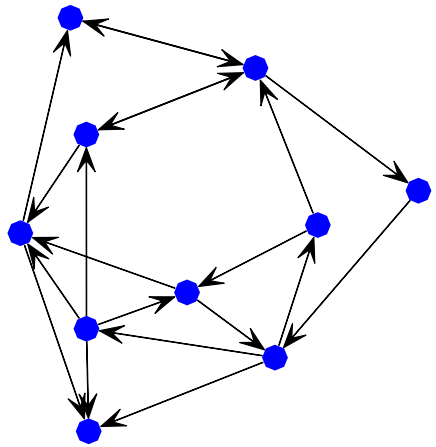
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



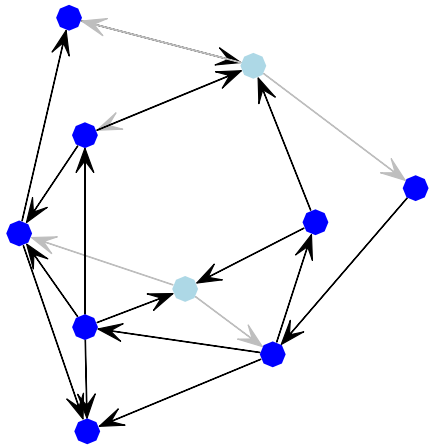
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



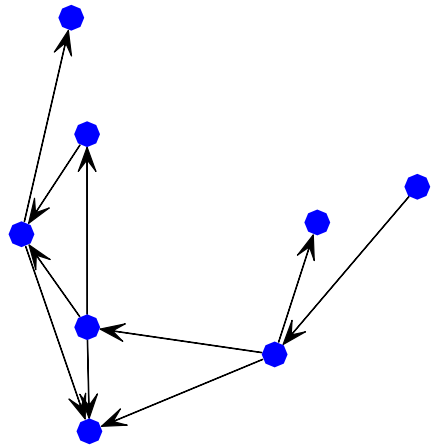
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



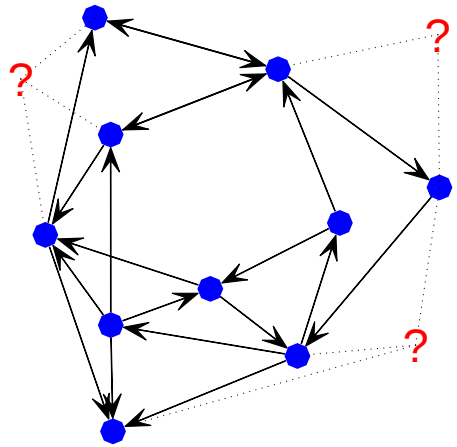
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

Frameworks for Statistical Analysis

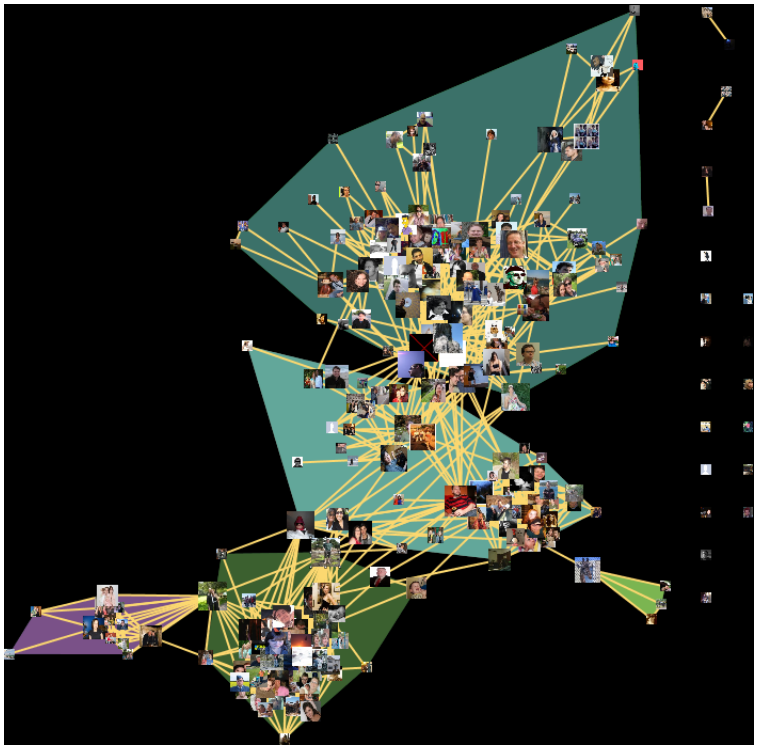
	Describe Structure	Describe Mechanism
Fully Observed Data	Description	Modeling (Statistical)
Partially Observed Data	Design-Based Inference	Likelihood Inference

Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

Descriptive Analysis

- Centrality
- Balance
- Reachability and path length
- Clustering (e.g. Facebook [image from Bernie Hogan, Oxford])



Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

The Horvitz-Thompson Estimator of a Total

I have a sample of students in a classroom and I want to estimate the total amount of money spent on textbooks in the class.

$$\hat{T}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

Where y_i is the amount spent by student i, and π_i is the probability i sampled.

$$E(\hat{T}_{HT}) = \sum_{i \in Pop} \frac{y_i}{\pi_i} \cdot \pi_i = \sum_{i \in Pop} y_i = T$$

$$Var(\hat{T}_{HT}) = \sum_{i \in Pop} \sum_{j \in Pop} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$

$$\hat{Var}(\hat{T}_{HT}) = \sum_{i \in s} \sum_{j \in s} \check{\Delta}_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Where $\check{\Delta}_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}$

Horvitz-Thompson Estimator for Number of Edges

How many friendships are in a classroom?

- I choose a simple random sample of n students, from a classroom of N students.
- I look at the web space of each student in the sample, and identify all other students in the class with whom they share friendships
- I use this data to form a Horvitz-Thompson estimator for the total number of friendships in the classroom

$$\hat{T}_{HT} = \sum_{i \in s} \sum_{j \notin s \text{ or } j > i} \frac{y_{ij}}{\pi_{ij}}$$

$$Var(\hat{T}_{HT}) = \sum_{i \in s} \sum_{j \notin s \text{ or } j > i} \sum_{k \in s} \sum_{l \notin s \text{ or } l > k} \Delta_{ij,kl} \frac{y_{ij}}{\pi_{ij}} \frac{y_{kl}}{\pi_{kl}}$$

Where $\Delta_{ij,kl} = \pi_{ij,kl} - \pi_{ij}\pi_{kl}$

The π_{ij}s

$$\pi_{ij} = 1 - \frac{\binom{N-2}{n}}{\binom{N}{n}} = \frac{n(2N-n-1)}{N(N-1)} \text{ for } i \neq j$$

$$\pi_{ij,kl} = p(\text{exactly one end of each sampled}) + p(\text{two end of one, one of the other sampled})$$
$$+ p(\text{both ends of both sampled})$$
$$= 2 \cdot 2\frac{\binom{N-4}{n-2}}{\binom{N}{n}} + 2 \cdot 2\frac{\binom{N-4}{n-3}}{\binom{N}{n}} + \frac{\binom{N-4}{n-4}}{\binom{N}{n}}$$

for i, j, k, l unique.

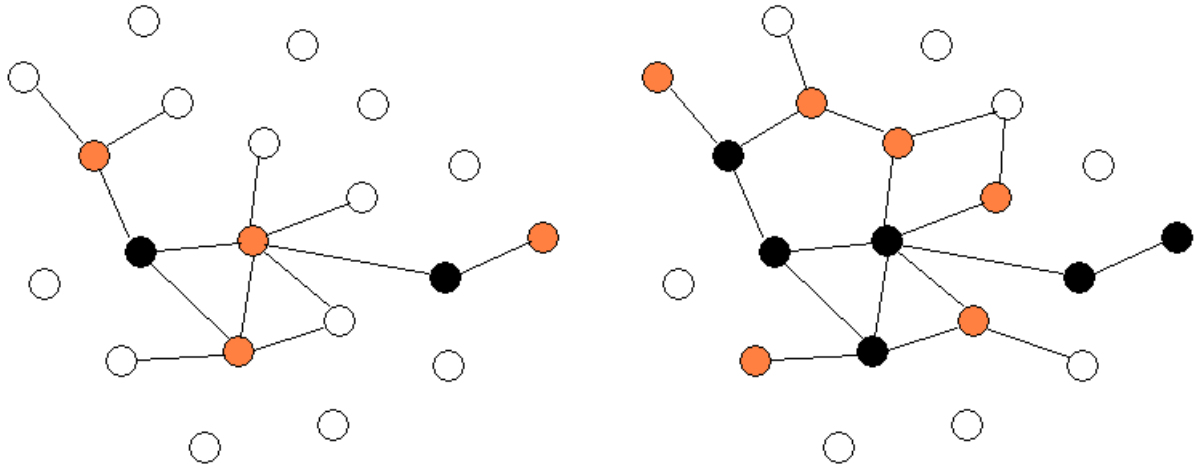
Link-Tracing Designs and HT Estimation

Consider:

$$\pi_i = 1 - \frac{\binom{N-m_i}{n}}{\binom{N}{n}}$$

Where m_i is the number of initial students that would have landed i in the sample.

1-Wave link-tracing: $m_i = 1+$ number of friends of i, observed!
2-Wave link-tracing: $m_i = 1+$ number of friends and friends of friends of i, NOT observed!



Limitation: Sampling probabilities often not obserable

Observable sampling probabilities under various sampling schemes for directed and undirected networks.

Sampling Scheme	Nodal Probabilities π_i		Dyadic Probabilities π_{ij}	
	Undirected	Directed	Undirected	Directed
Ego-centric	X	X	X	X
One-Wave	X			
$k-$Wave, $1 < k < \infty$				
Saturated	X			

- Nodal and dyadic sampling probabilities are considered separately.
- "X" indicates observable sampling probabilities, while a blank indicates unobservable sampling probabilities.

Frameworks for Statistical Analysis

	Describe Structure	Describe Mechanism
Fully Observed Data	Description	Modeling (Statistical)
Partially Observed Data	Design-Based Inference	Likelihood Inference

Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

Modeling Social Network Data - Independence Models

Y =

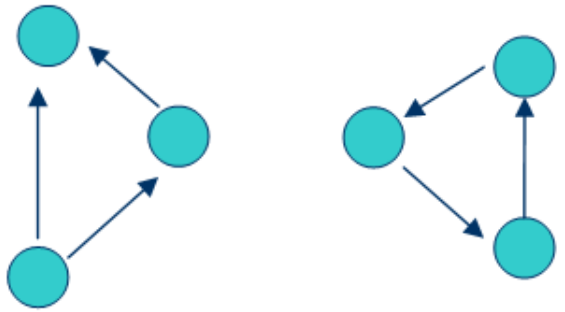
	A	B	C	D
A	-	1	0	0
B	0	-	1	1
C	0	0	-	0
D	1	1	1	-

$$logitP(Y_{ij} = 1|X) = \beta_0 + \beta_1(\text{same sex}) + \beta_2(\text{i older than j}) + ..$$

Logistic regression form. Each pair (dyad) independent of all others.

Scientific questions answered by comparing estimated coefficients, β_i, to 0.

More complex models: Exponential-family Random Graph Models



- Sometimes, we are interested in more complex relational structures.
- Need to look at whole graph at once, can't separate dyads.

Exponential-family Random Graph Model (ERGM):

(Holland and Leinhardt (1981), Snijders et al, (2006),...)

$$P_\beta(Y = y) = c(\beta)e^{\beta_1 g_1(y)+\beta_2 g_2(y)+\ldots}$$

- $g(y)$ represent components of the social process
- $c(\beta)$ is the normalizing constant

Fitting Models to Partially Observed Social Network Data

- Two types of data: Observed relations (Y_{obs}), and indicators of units sampled (D).

$$P(Y_{obs}, D|\beta, \delta) = \sum_{Unobserved} P(Y, D|\beta, \delta)$$
$$= \sum_{Unobserved} P(D|Y, \delta)P(Y|\beta)$$

- β is the model parameter
- δ is the sampling parameter

If $P(D|Y, \delta) = P(D|Y_{obs}, \delta)$ (*adaptive sampling* or *missing at random*)

Then

$$P(Y_{obs}, D|\beta, \delta) = P(D|Y, \delta) \sum_{Unobserved} P(Y|\beta)$$

- Can find maximum likelihood estimates by summing over the possible values of unobserved, ignoring sampling
- Sample with Markov Chain Monte Carlo (MCMC)

Fitting Models to Partially Observed Social Network Data

- Two types of data: Observed relations (Y_{obs}), and indicators of units sampled (D).

$$P(Y_{obs}, D|\beta, \delta) = \sum_{Unobserved} P(Y, D|\beta, \delta)$$
$$= \sum_{Unobserved} P(D|Y, \delta)P(Y|\beta)$$

- β is the model parameter
- δ is the sampling parameter

If $P(D|Y, \delta) = P(D|Y_{obs}, \delta)$ (*adaptive sampling* or *missing at random*)

Then

$$P(Y_{obs}, D|\beta, \delta) = P(D|Y, \delta) \sum_{Unobserved} P(Y|\beta)$$

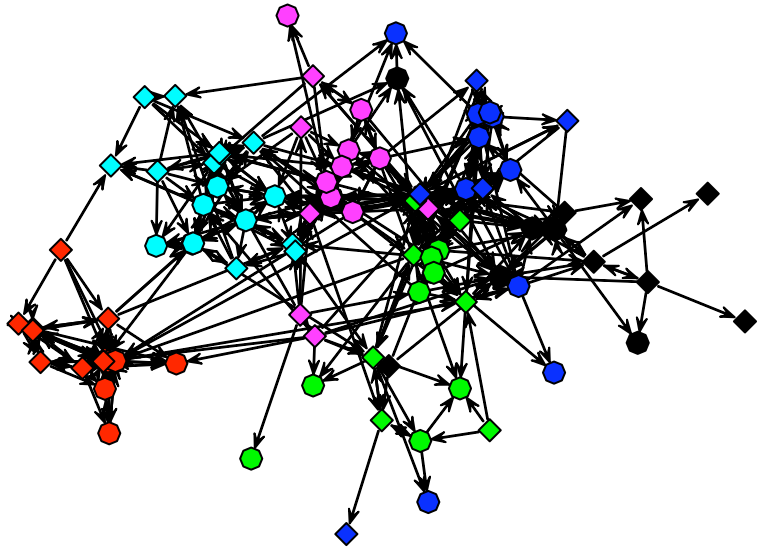
- Can find maximum likelihood estimates by summing over the possible values of unobserved, ignoring sampling
- Sample with Markov Chain Monte Carlo (MCMC)

Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

Example: Friendships in a School

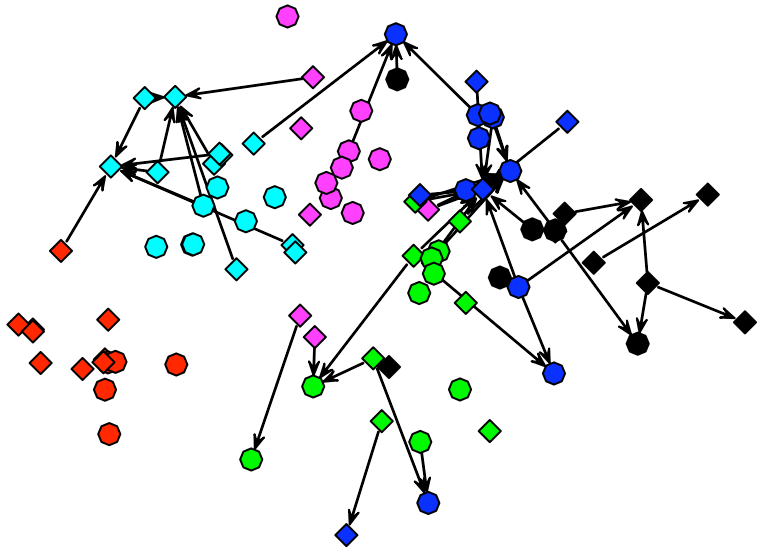
From the National Longitudinal Survey on Adolescent Health - Wave 1:



- Each student asked to nominate up to 5 male and 5 female friends
- Sex and Grade available for 89 students, 70 students reported friendships.

Example: Friendships in a School

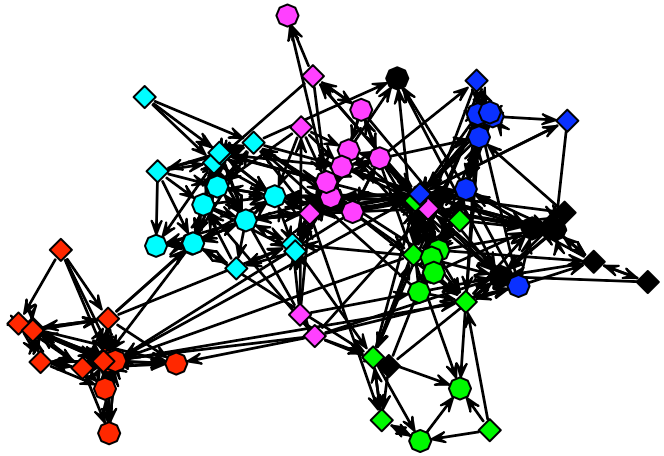
From the National Longitudinal Survey on Adolescent Health - Wave 1:



- Each student asked to nominate up to 5 male and 5 female friends
- Sex and Grade available for 89 students, 70 students reported friendships.

Example: Friendships in a School

From the National Longitudinal Survey on Adolescent Health - Wave 1:



- Each student asked to nominate up to 5 male and 5 female friends
- Sex and Grade available for 89 students, 70 students reported friendships.

Example: Friendships in a School

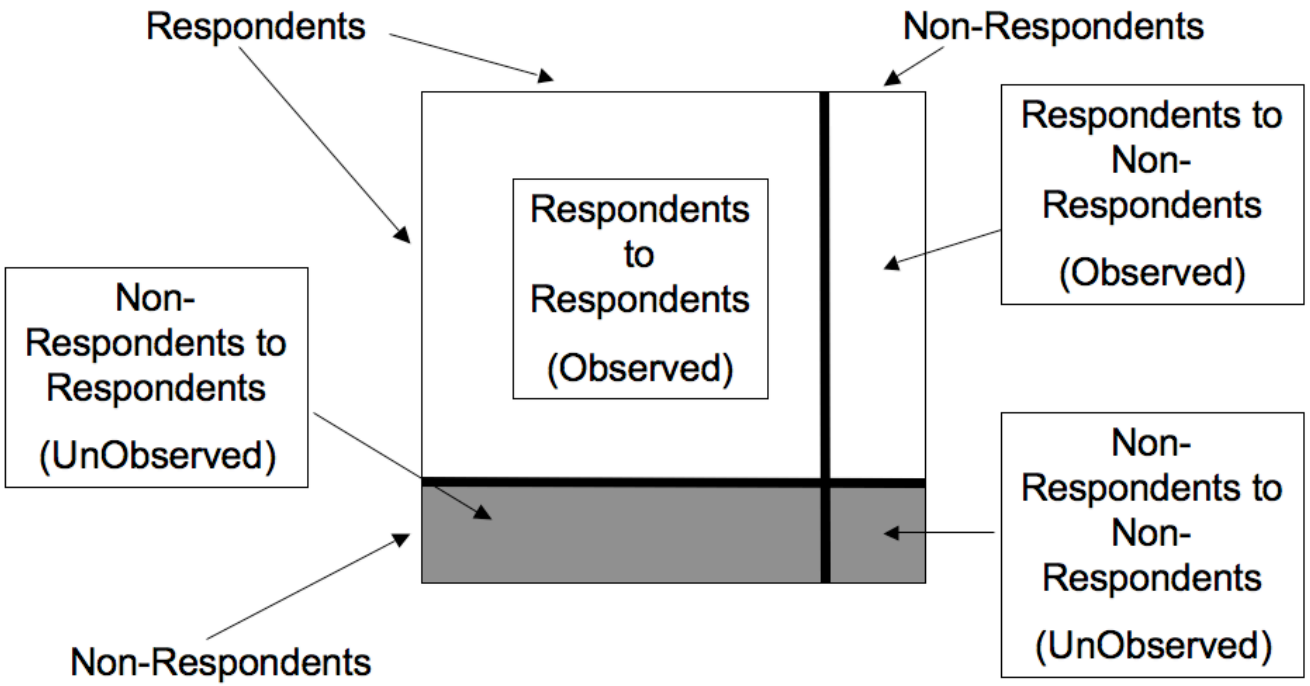
- **Scientific Question:** Do friendships form in an egalitarian or an hierarchal manner?
- **Methodological Question:** Can we fit a network model to a network with missing data? Is the fit different from that of just the observed data?

$P(D|Y,\delta) = P(D|Y_{obs},\delta)$ (*missing at random*)

Does observed status depend on unobserved characteristics?

Structure of Data

- Up to 5 female friends and up to 5 male friends
- 89 students in school
- 70 completed friendship nominations portion of survey



Example: Friendships in a School

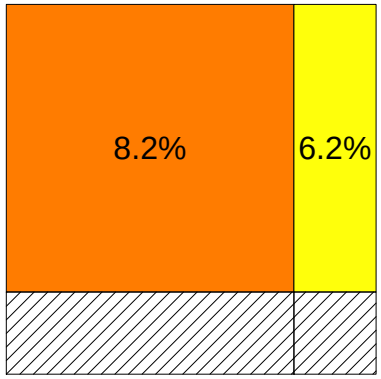
Fit an ERGM to the partially observed data, get coefficients like in logistic regression.

Terms in the model:

- **Density**: Overall rate of ties
- **Reciprocity**: Do students tend to reciprocate nominations?
- **Popularity by Grade**: Do students in different grades receive different rates of ties?
- **Popularity by Sex**: Do boys and girls receive different rates of ties?
- **Age:Sex Mixing**: Rates of ties between older and younger boys and girls
- Propensity for ties within sex and grade to be **transitive** (hierarchal)
- Propensity for ties within sex and grade to be **cyclical** (egalitarian)
- **Isolation**: Propensity for students to receive no nominations

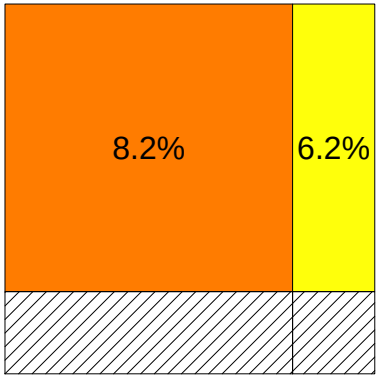
Percent of Possible Relations Realised

	Observed
Respondents to Respondents	8.2
Respondents to Non-Respondents	6.2
Non-Respondents to Respondents	-
Non-Respondents to Non-Respondents	-

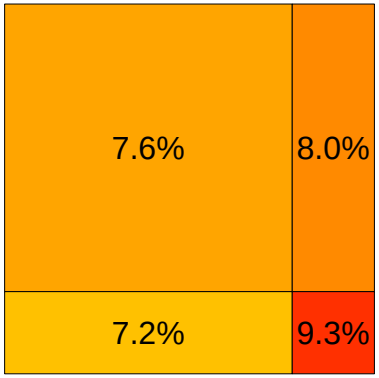


Goodness of Fit: Percent of Possible Relations Realised

	Observed	Fit
Respondents to Respondents	8.2	7.6
Respondents to Non-Respondents	6.2	8.0
Non-Respondents to Respondents	-	7.2
Non-Respondents to Non-Respondents	-	9.3



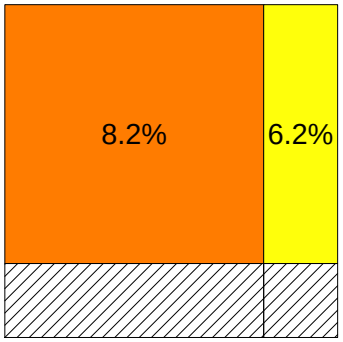
(e) Observed



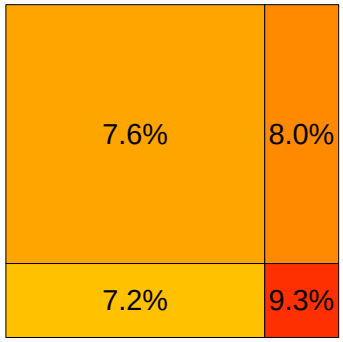
(f) Fit

Goodness of Fit: Percent of Possible Relations Realised

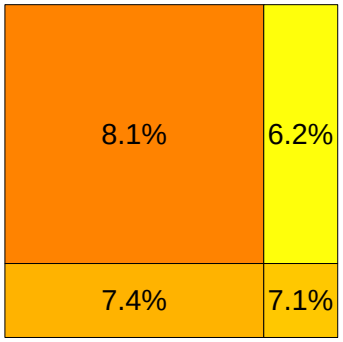
	Observed	Original	Diff. Popularity
Respondents to Respondents	8.2	7.6	8.1
Respondents to Non-Respondents	6.2	8.0	6.2
Non-Respondents to Respondents	-	7.2	7.4
Non-Respondents to Non-Respondents	-	9.3	7.1



(g) Observed



(h) Original



(i) Differential Popularity

	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

Conclusions, School Friendships Example

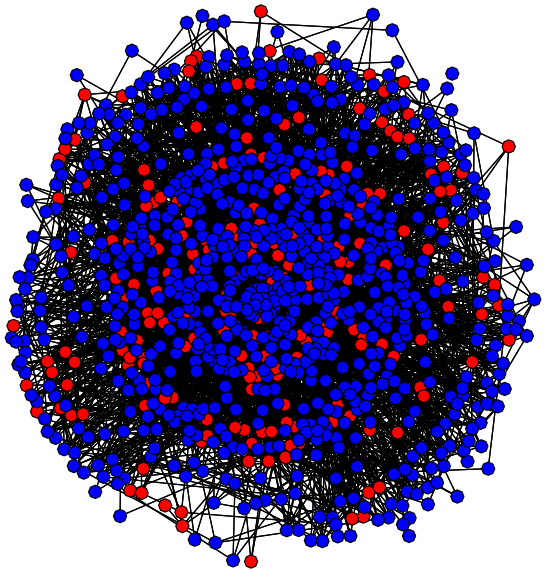
- Nominations are reciprocated at a higher rate than random
- Males receive nominations from other males at a higher rate than females from females
- Nominations within grade are more likely than outside grade
- Nominations of older students are more likely than younger students
- Nominations within sex and grade are more consistent with a hierarchal rather than egalitarian structure
- More students receive no nominations than we would expect at random.

Frameworks for Statistical Analysis

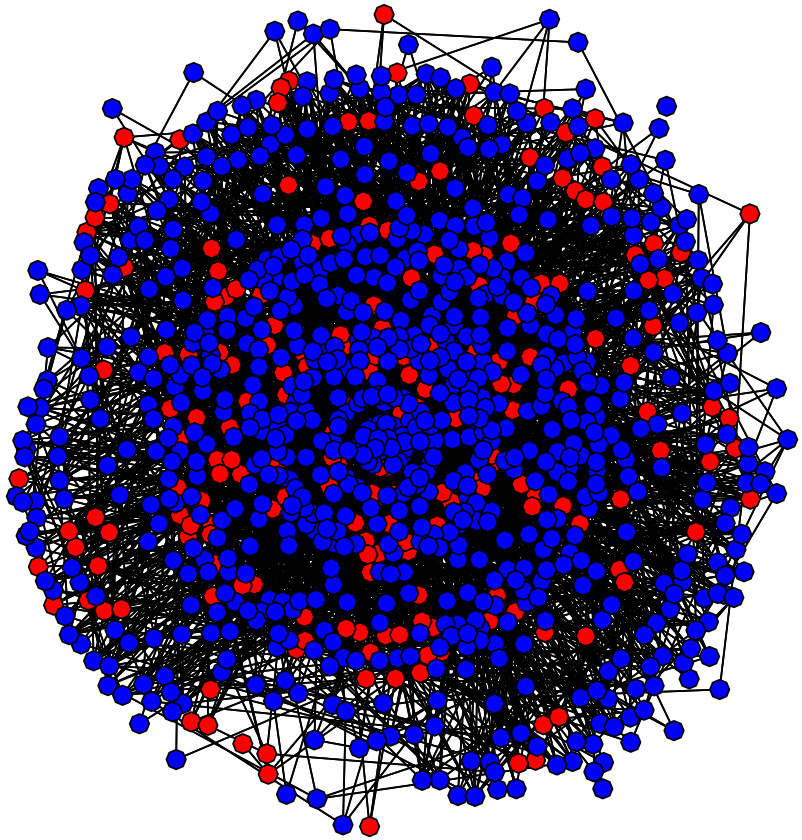
	Describe Structure	Describe Mechanism
Fully Observed Data	Description	Modeling (Statistical)
Partially Observed Data	Design-Based Inference	Likelihood Inference

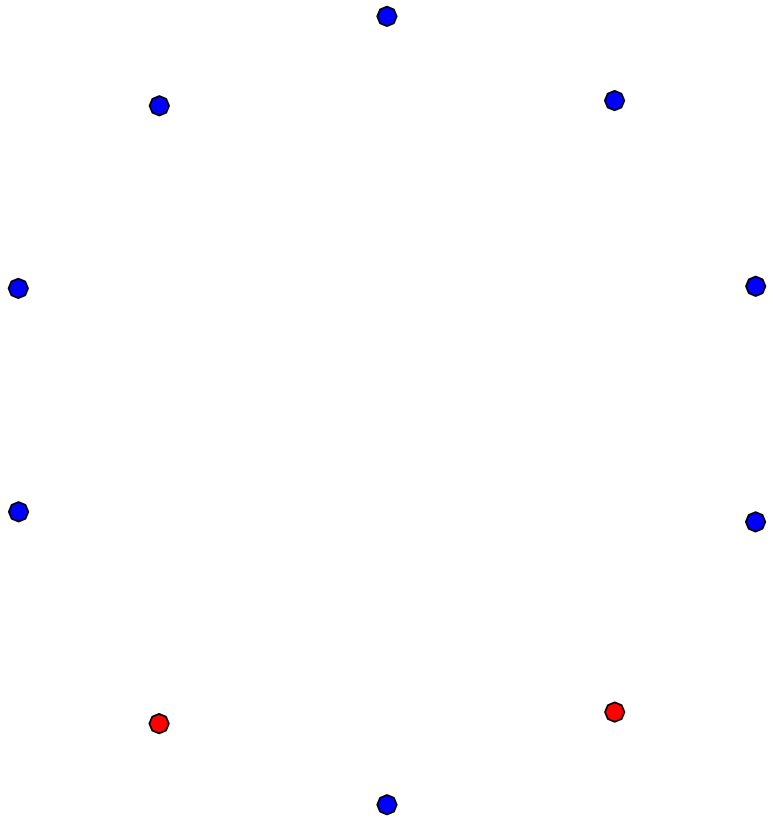
Injection Drug User Example

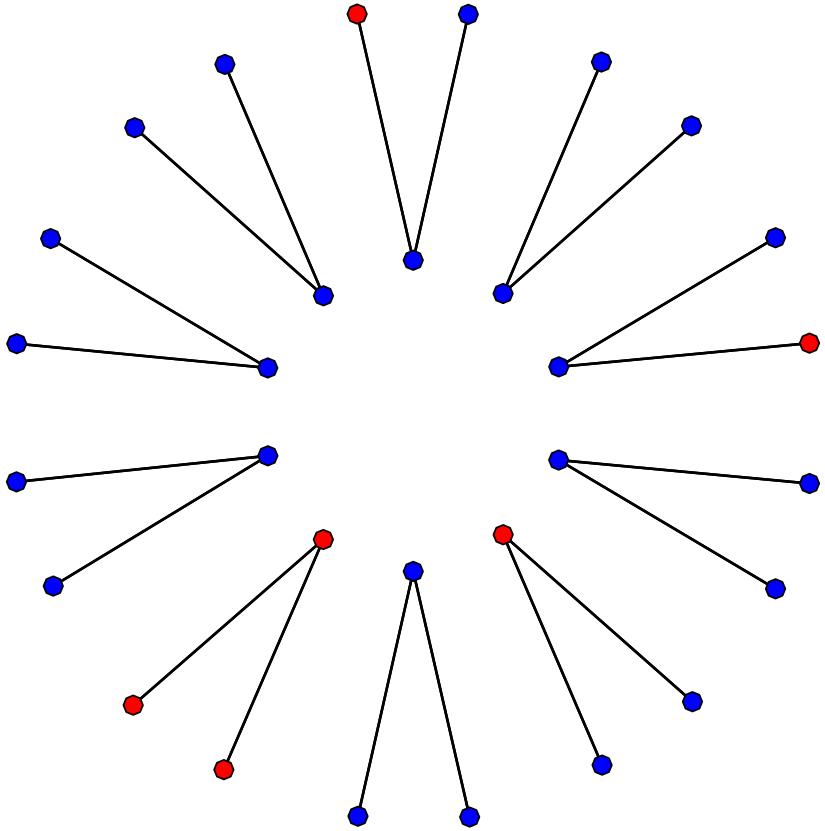
Simulations corresponding to U.S. Center for Disease Control study of Injection Drug Users in New York City

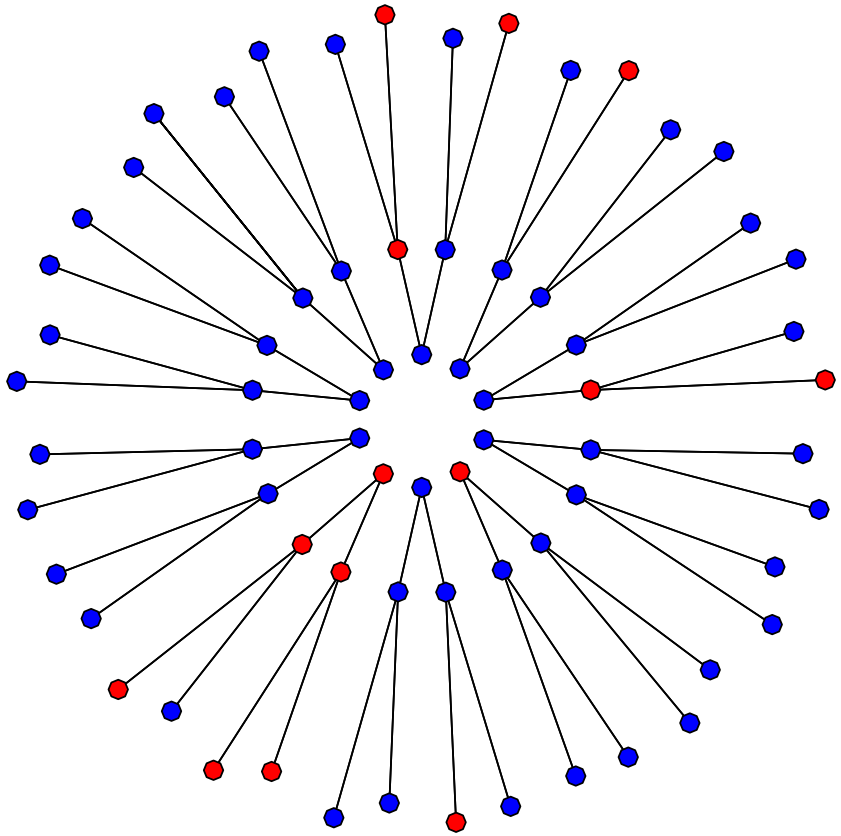


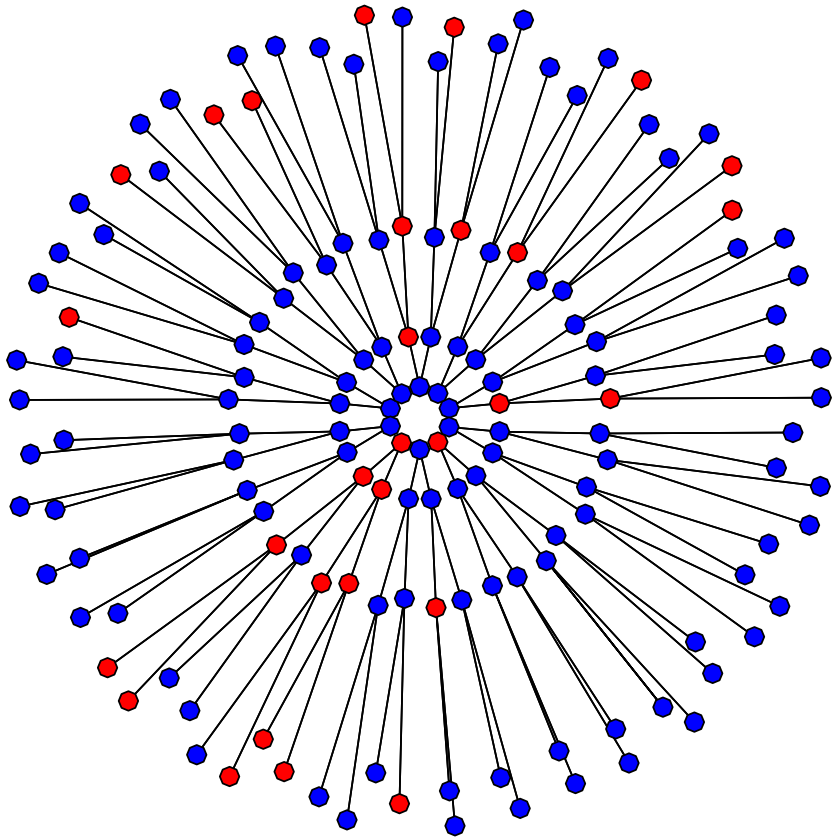
- Injection Drug Users (IDU) asked how many IDUs they know, and HIV tested.
- Given 2 coupons to pass to other IDUs they know, to invite them to join study.
- Sampling continues until sample size 500 (from simulated population 715).

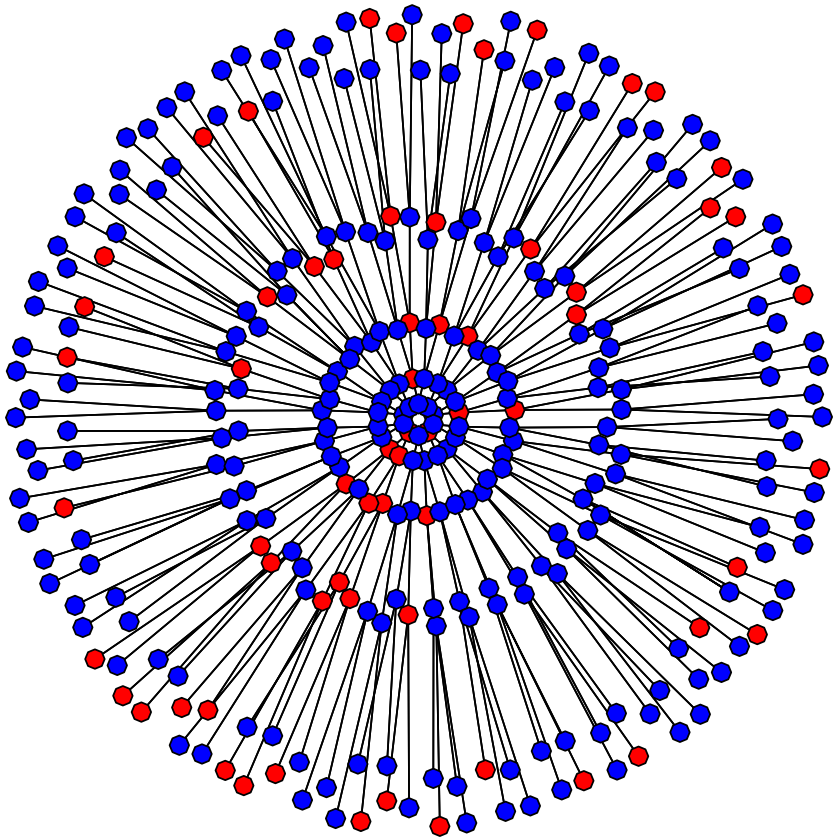


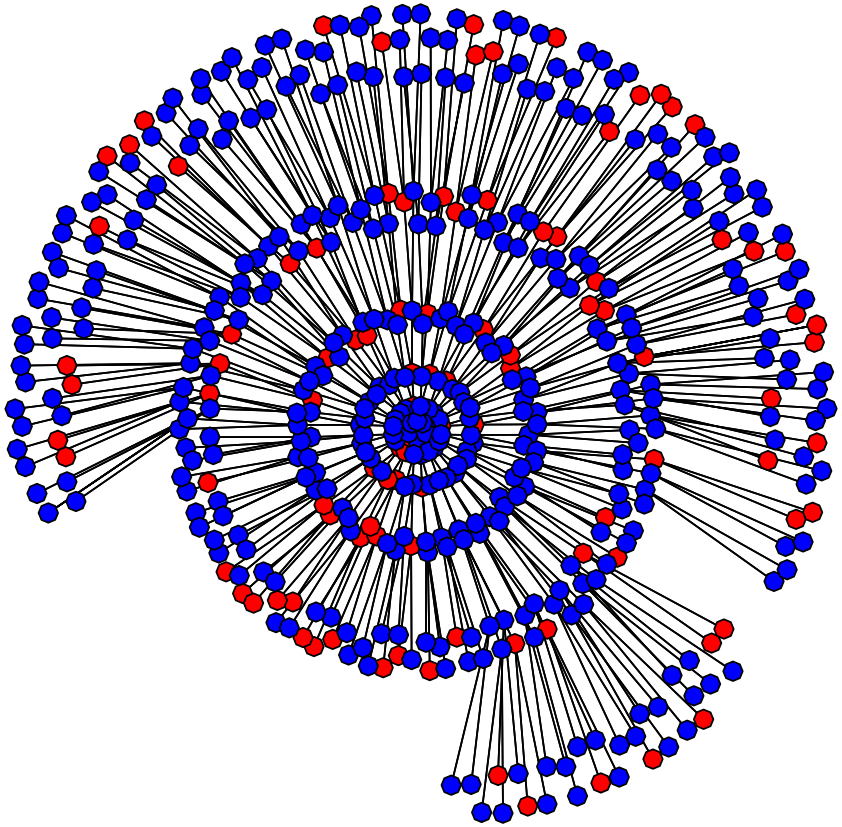


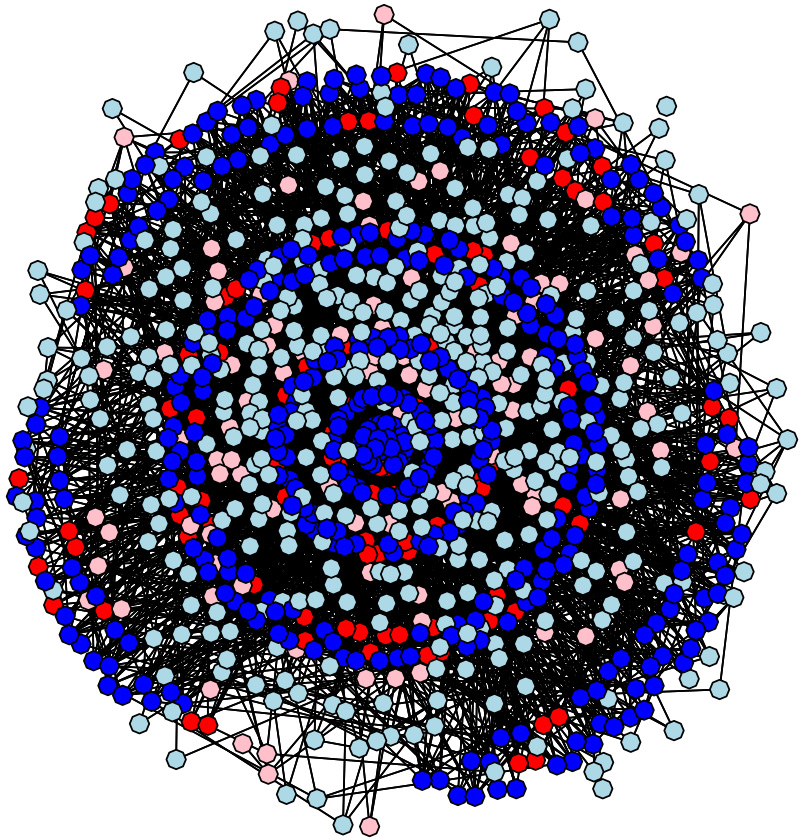












Injection Drug User Example

- **Scientific Question:** What proportion of Injection Drug Users in New York City are HIV positive?
- **Methodological Question:** Can we estimate population proportions from samples starting at a convenience sample of seeds?

$P(D|Y,\delta) = P(D|Y_{obs},\delta)$ (*adaptive sampling*)

$$P(D|Y,\delta) = P(seeds|Y,\delta)P(D|Y,\delta,seeds) = P(seeds)P(D|Y_{obs},\delta,seeds)$$

but typically $P(seeds|Y,\delta) \neq P(seeds|Y_{obs},\delta)$

Structure of Analysis

Sample:

- Link-tracing sampling variant - *Respondent-Driven Sampling*
- Ask number of contacts - but not who. Can't identify alters. No matrix.
- Network used as sampling tool

Existing Approach:

- Assume inclusion probability proportional to number of contacts (Volz and Heckathorn, 2008)
- Assume many waves of sampling remove bias of seed selection

Our work:

- Design-based (describe structure, not mechanism)
- Fit simple network model to observed data (model-assisted)
- Correct for biases due to network-based sampling, and observable irregularities

Generalized Horvitz-Thompson Estimator

- Goal: Estimate proportion "infected" :

$$\mu = \frac{1}{N}\sum_{j=1}^{N} z_i$$

where

$$z_i = \begin{cases} 1 & \text{node } i \text{ infected} \\ 0 & \text{node } i \text{ uninfected.} \end{cases}$$

- Generalized Horvitz-Thompson Estimator:

$$\hat{\mu} = \frac{\sum_{i:S_i=1} \frac{1}{\pi_i} z_i}{\sum_{i:S_i=1} \frac{1}{\pi_i}}$$

where

$$S_i = \begin{cases} 1 & \text{node } i \text{ sampled} \\ 0 & \text{node } i \text{ not sampled} \end{cases} \qquad \pi_i = P(S_i = 1).$$

Key Point: Requires $\pi_i \, \forall \, i : S_i = 1$

Simulation Study

Simulate Population

- 1000, 835, 715, 625, 555, or 525 nodes
- 20% "Infected"

Simulate Social Network (from ERGM, using `statnet`)

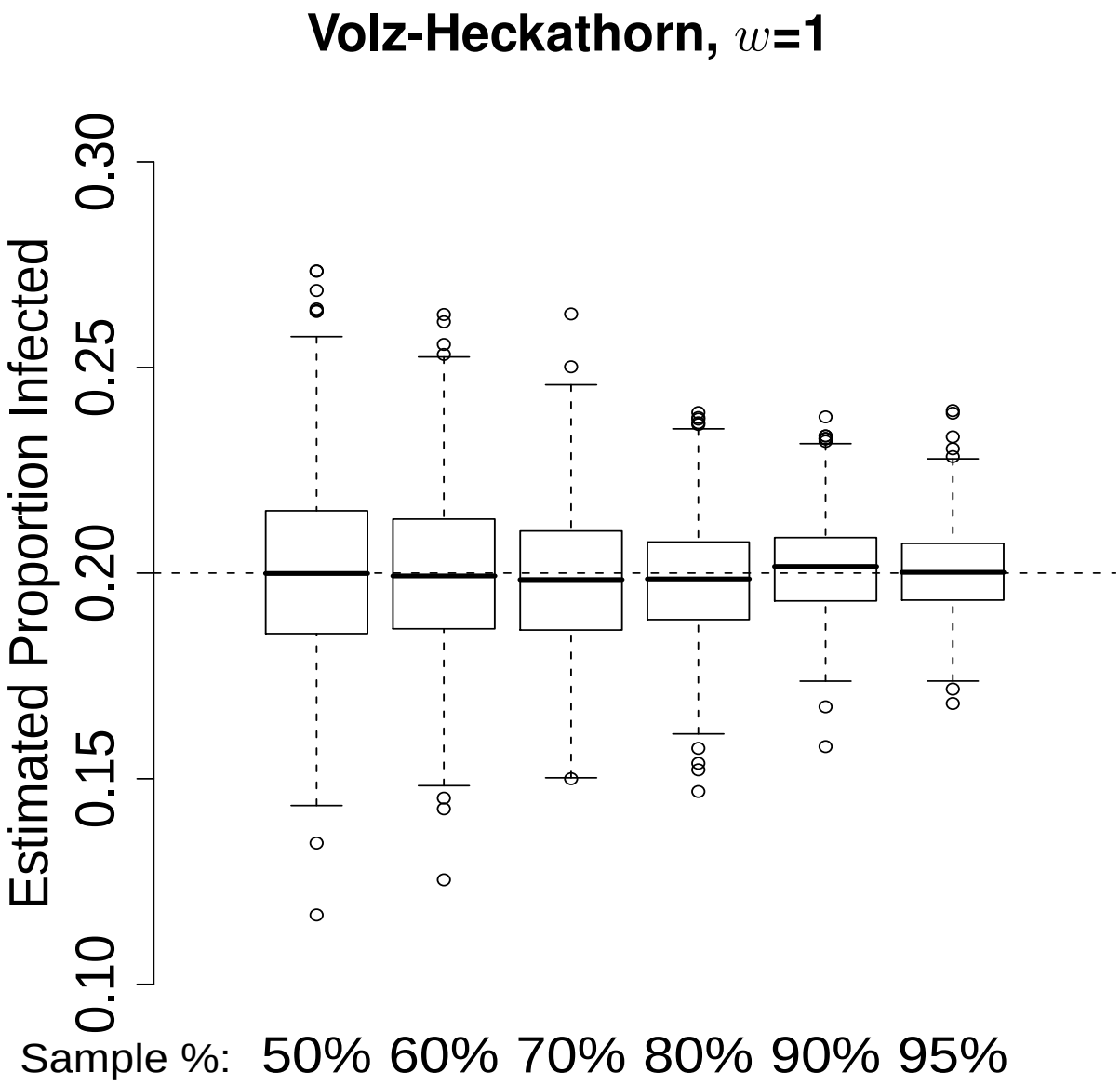
- Mean degree 7
- Homophily on Infection: $R = \frac{P(\text{infected to infected tie})}{P(\text{uninfected to infected tie})} = 5$ (or other)
- Differential Activity: $w = \frac{\text{mean degree infected}}{\text{mean degree uninfected}} = 1$ (or other)

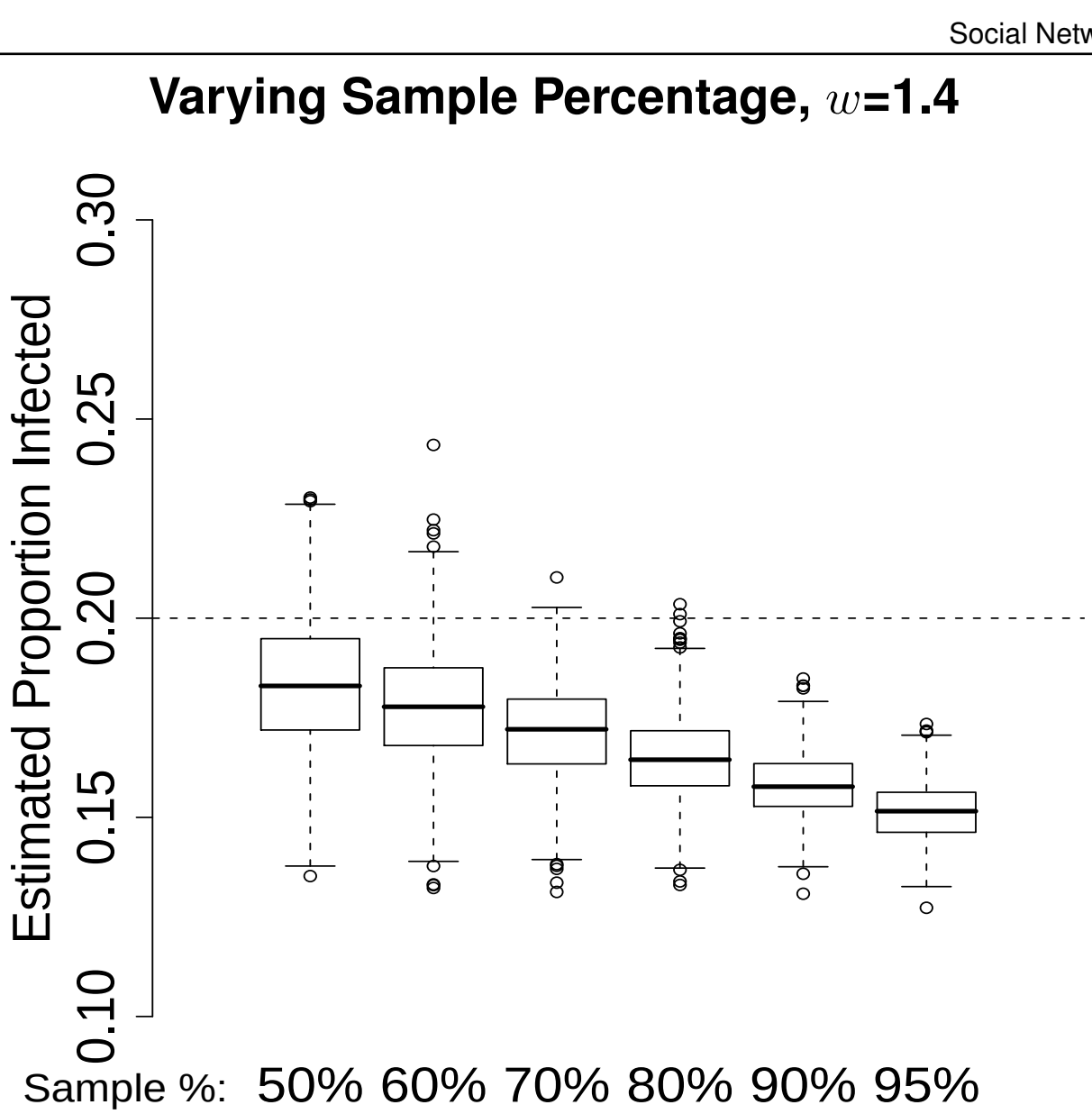
Simulate Respondent-Driven Sample

- 500 total samples
- 10 seeds, chosen proportional to degree
- 2 coupons each
- Coupons at random to relations
- Sample without replacement

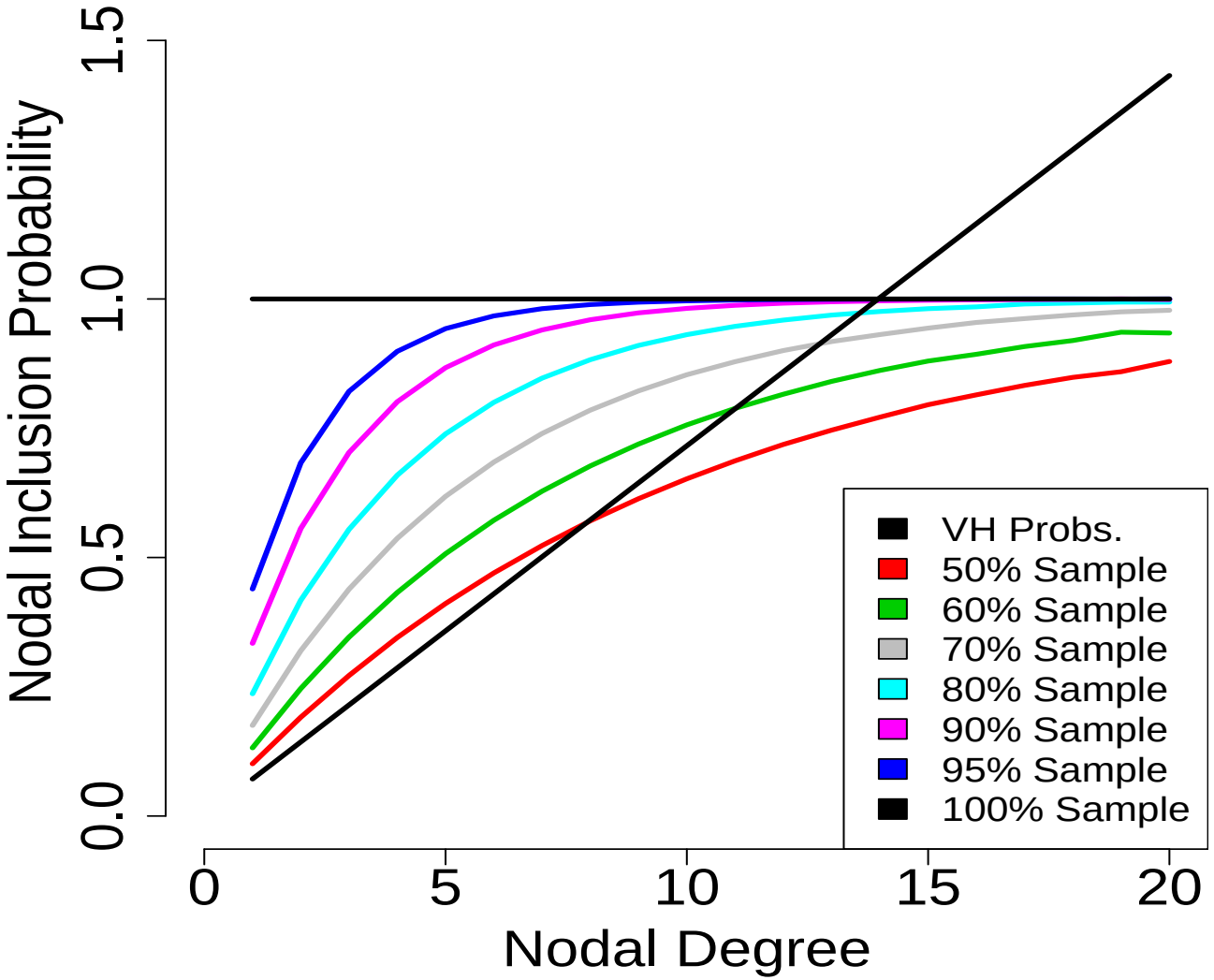
Repeat 1000 times!

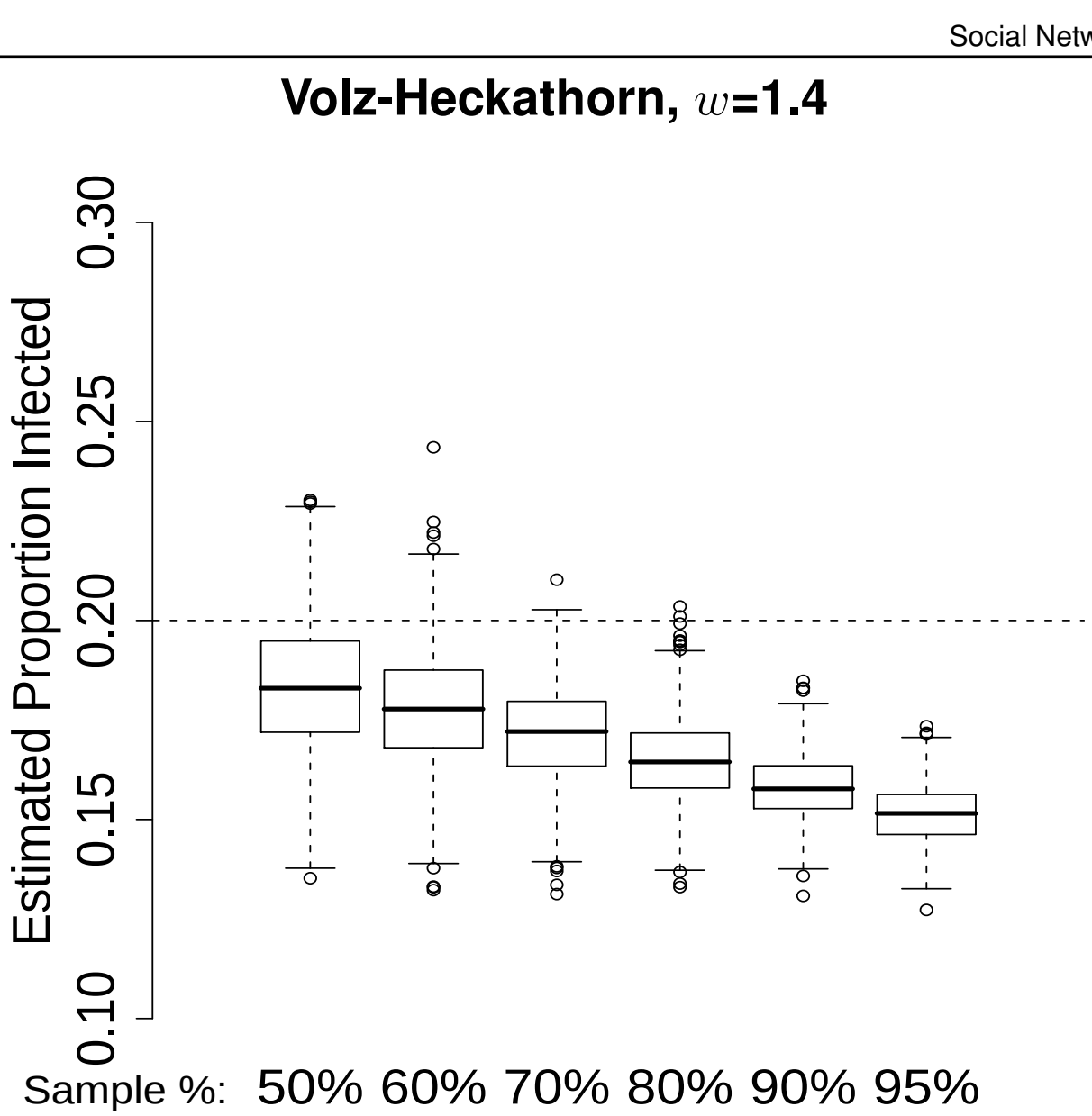
Blue parameters varied in study.

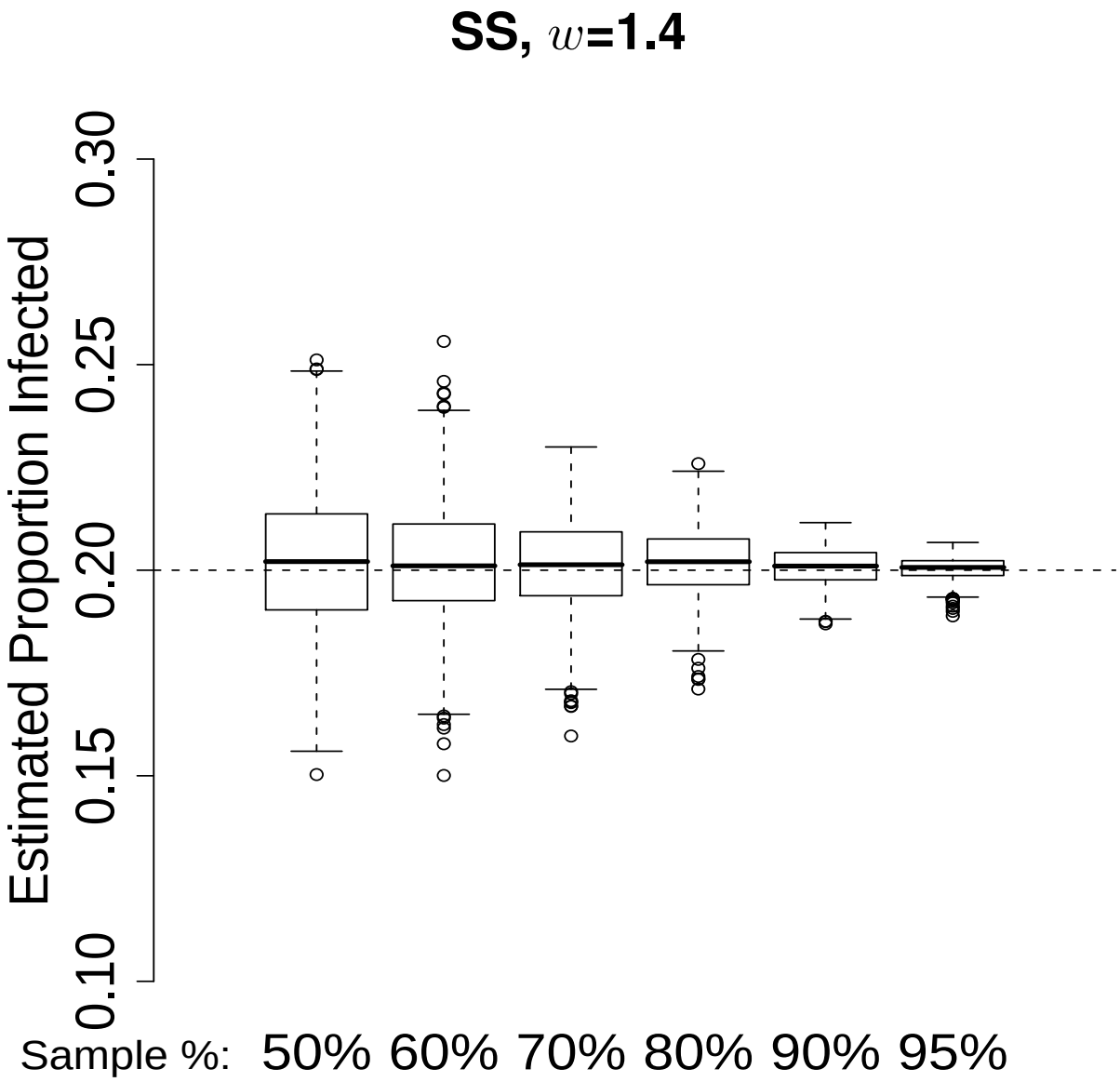

Volz-Heckathorn, w=1
Estimated Proportion Infected
0.10
0.15
0.20
0.25
0.30
Sample %: 50% 60% 70% 80% 90% 95%


Varying Sample Percentage, w=1.4
Estimated Proportion Infected
0.10
0.15
0.20
0.25
0.30
Sample %: 50% 60% 70% 80% 90% 95%

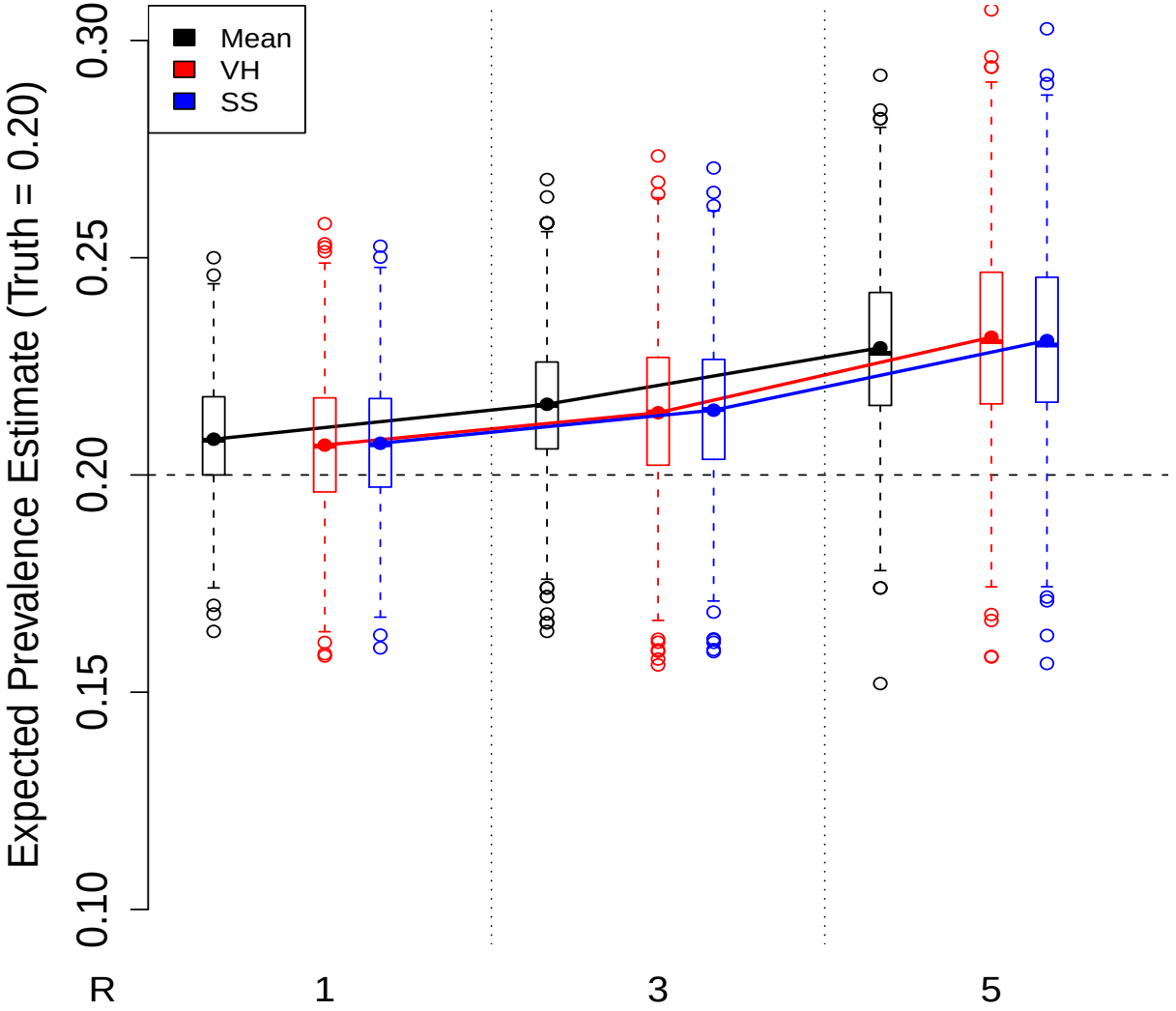
Successive Sampling Mapping



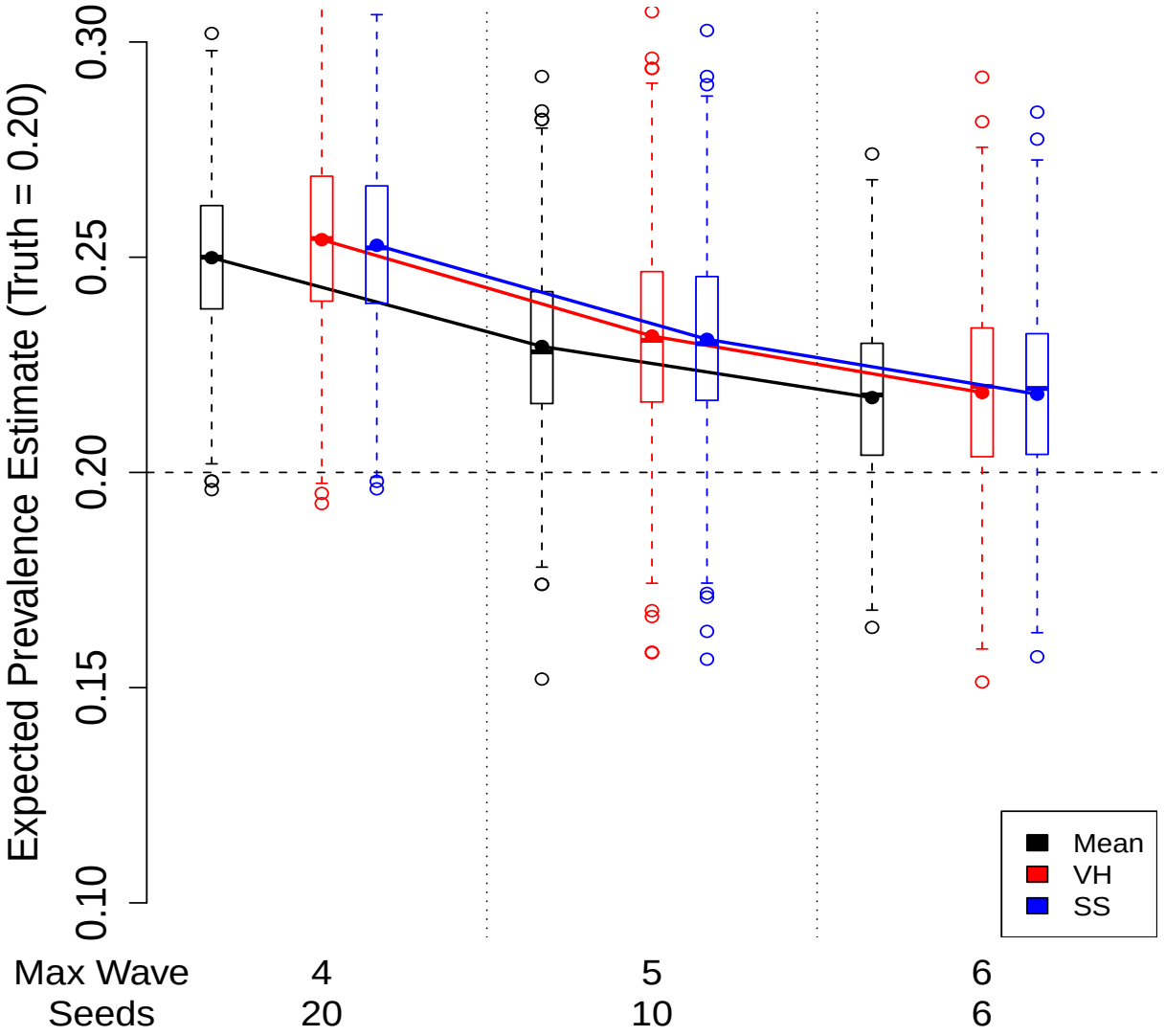

Volz-Heckathorn, w=1.4
Estimated Proportion Infected
0.10
0.15
0.20
0.25
0.30
Sample %: 50% 60% 70% 80% 90% 95%


SS, w=1.4
Estimated Proportion Infected
0.10
0.15
0.20
0.25
0.30
Sample %: 50% 60% 70% 80% 90% 95%

All Infected Seeds, varying Homophily, 50%



All Infected Seeds, varying number of seeds, 50%



Network-Model Estimator

Key Points:

- Goal: Estimate inclusion probabilities to use to weight sampled nodes
- If we knew the network, we could simulate sampling to estimate inclusion probabilities

Approach:

1. Use (weighted) information in the data to estimate network structure
2. Use simulated sampling over estimated network to estimate inclusion probabilities
3. Iterate steps (1) and (2)

Simulation Study

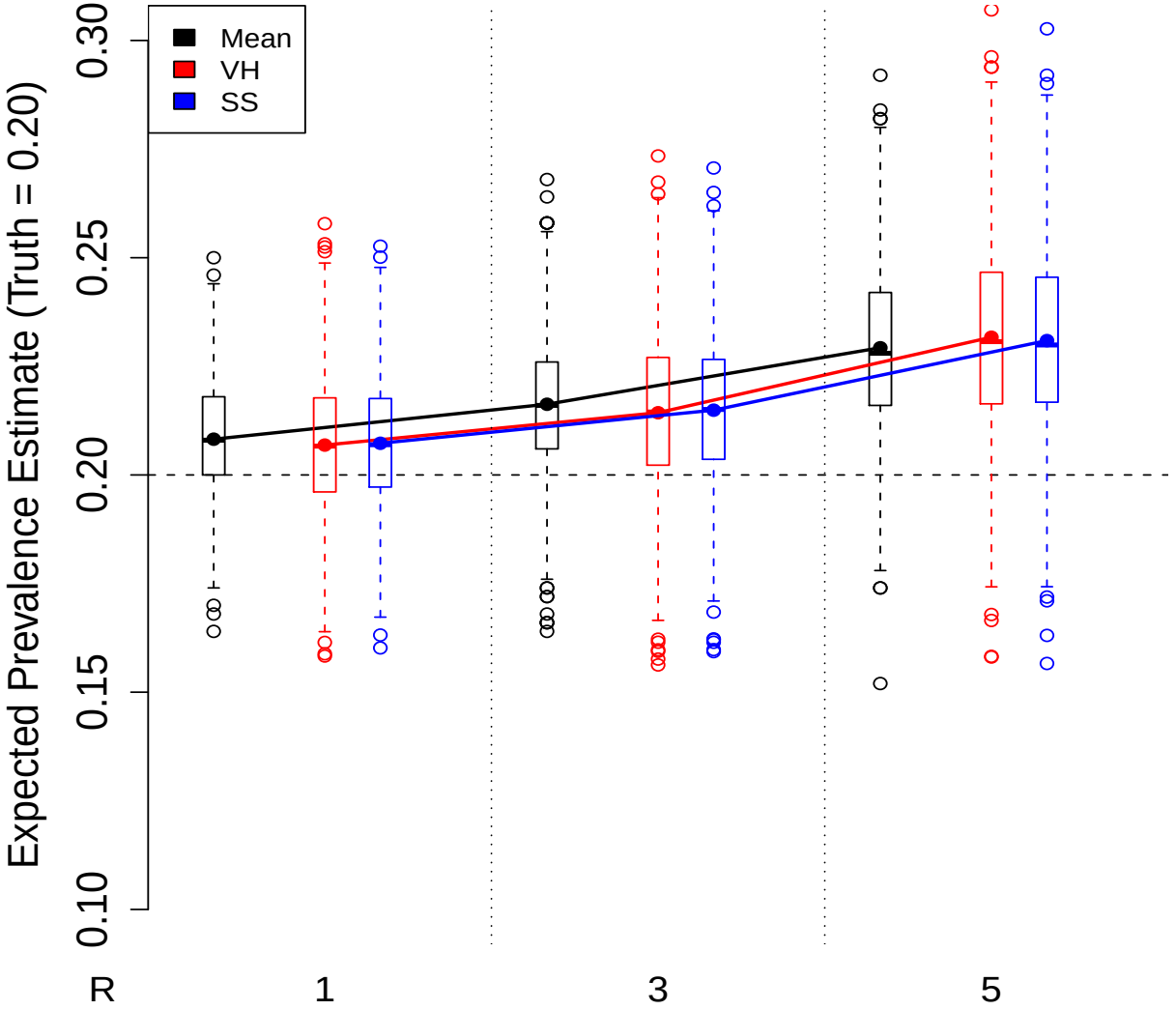
Set-up:

- High sample fraction (500 of 715)
- Infected have more relations than uninfected (twice as many)
- Initial sample biased (all infected)
- Truth: 20% Infected

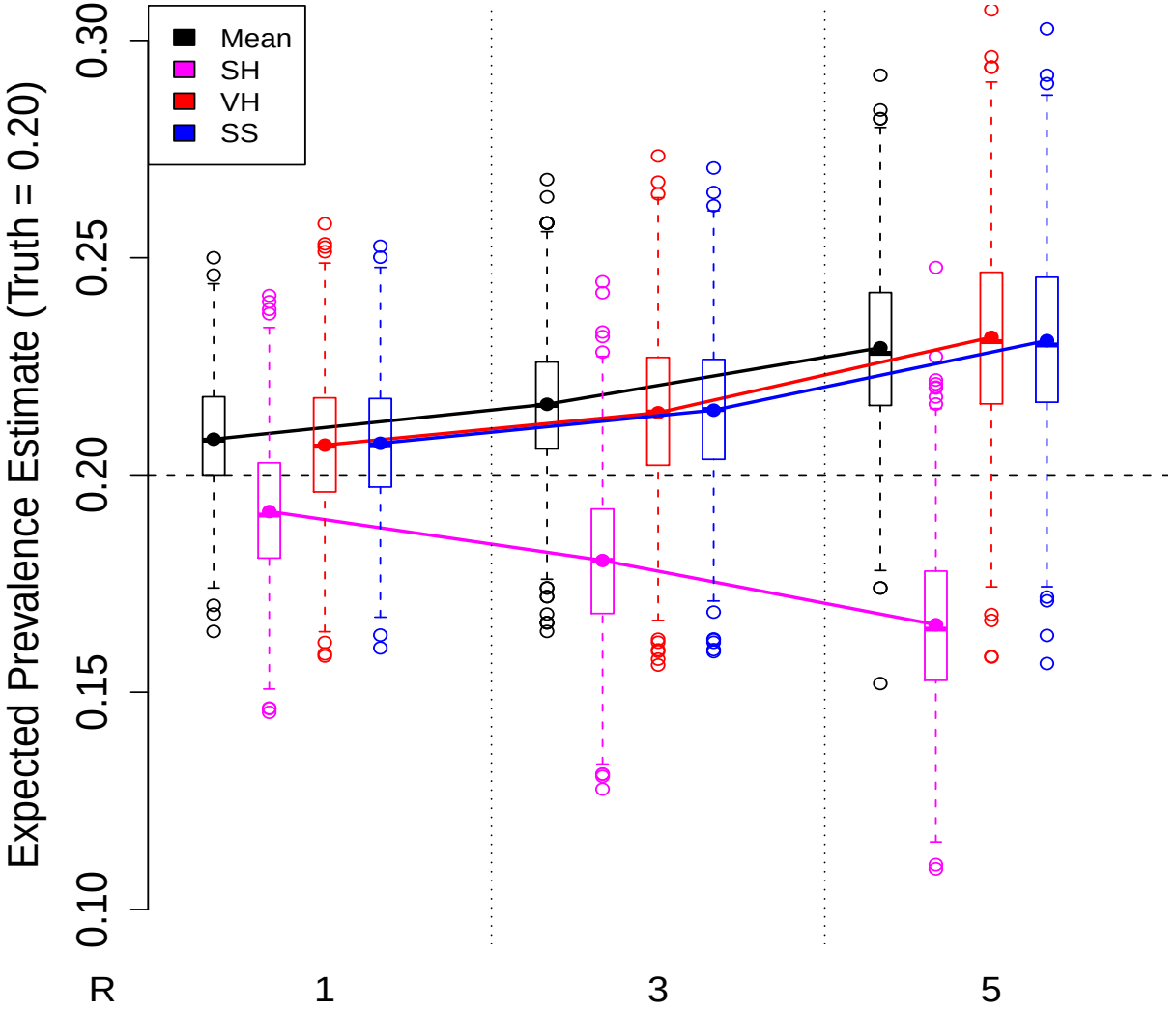
Comparison of Estimators:

- "Mean" : Naive Sample Mean
- SH: Salganik-Heckathorn: based on MME of number of cross-relations
- VH: Existing Volz-Heckathorn Estimator (2008)
- SS: Another new estimator, not discussed here
- Net: New Network-Based Estimator

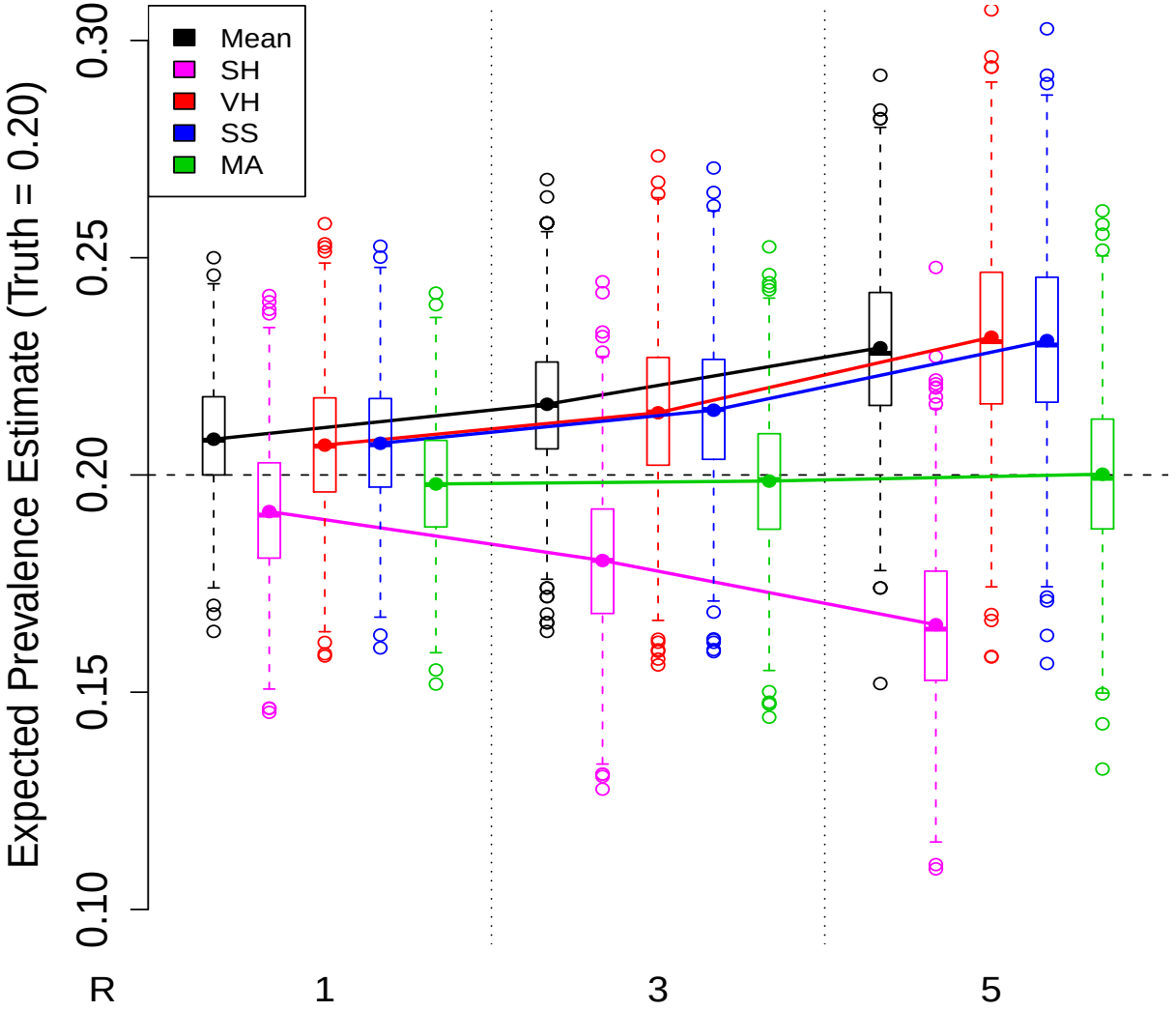
All Infected Seeds, varying Homophily



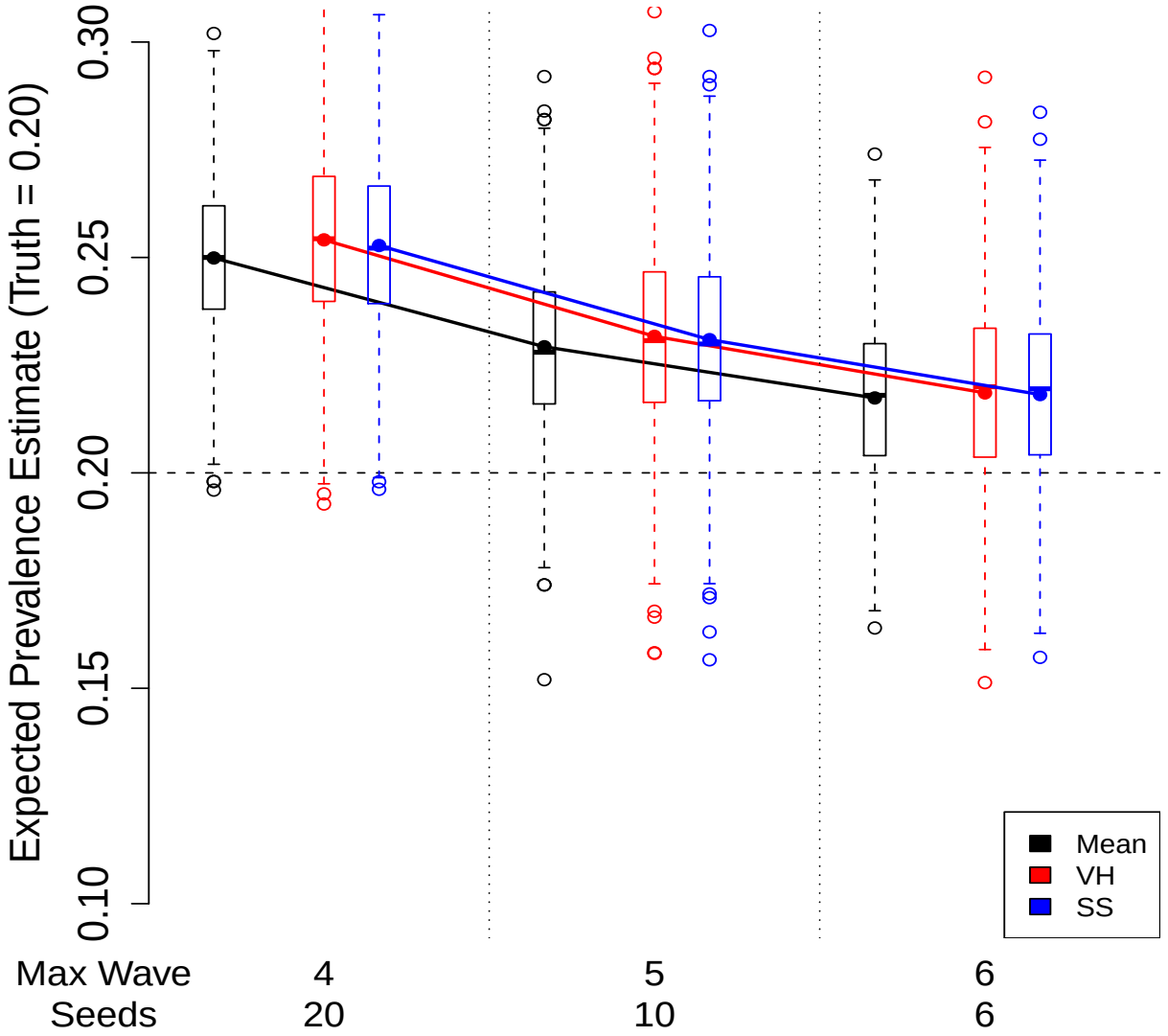
All Infected Seeds, varying Homophily



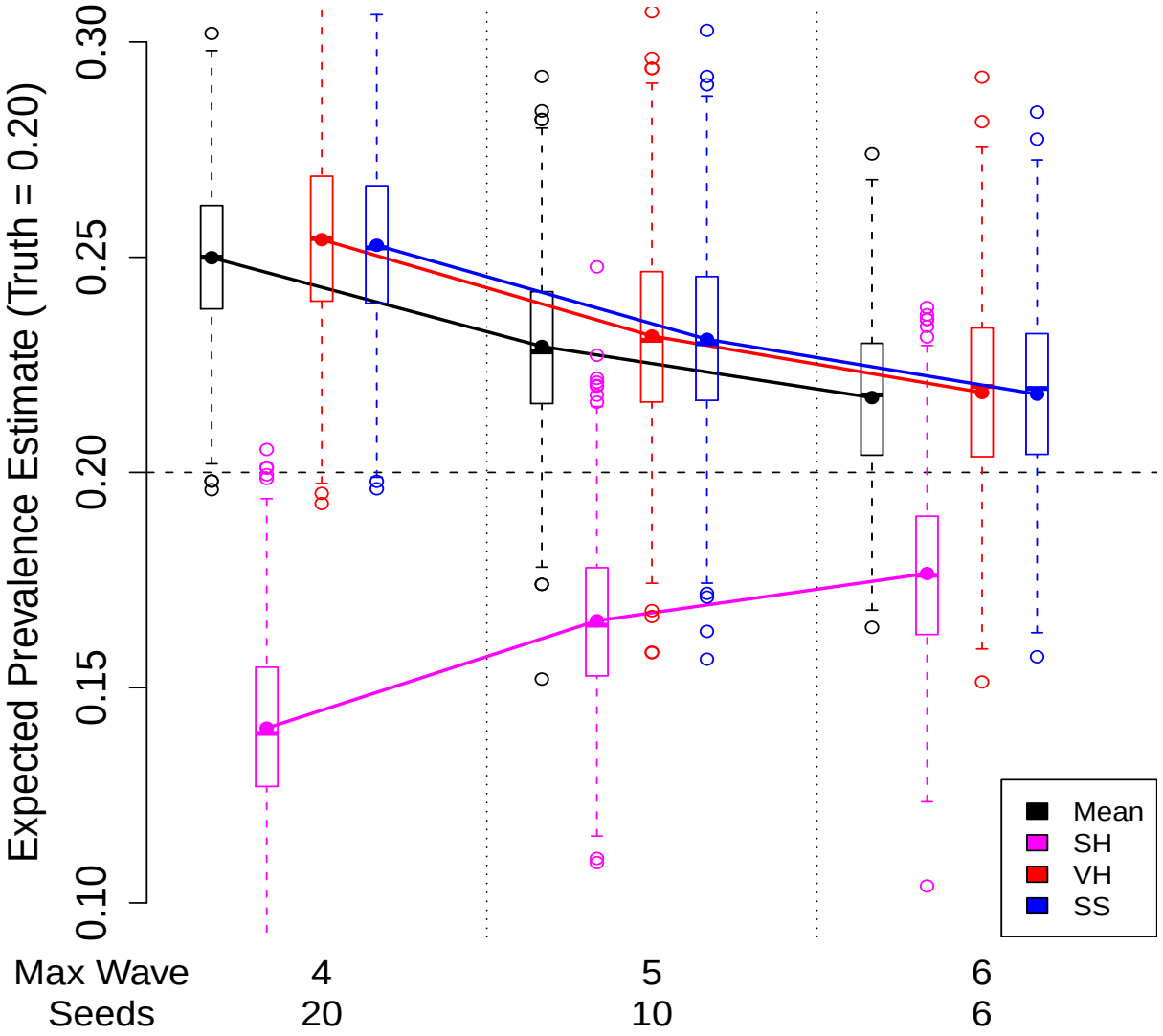
All Infected Seeds, varying Homophily



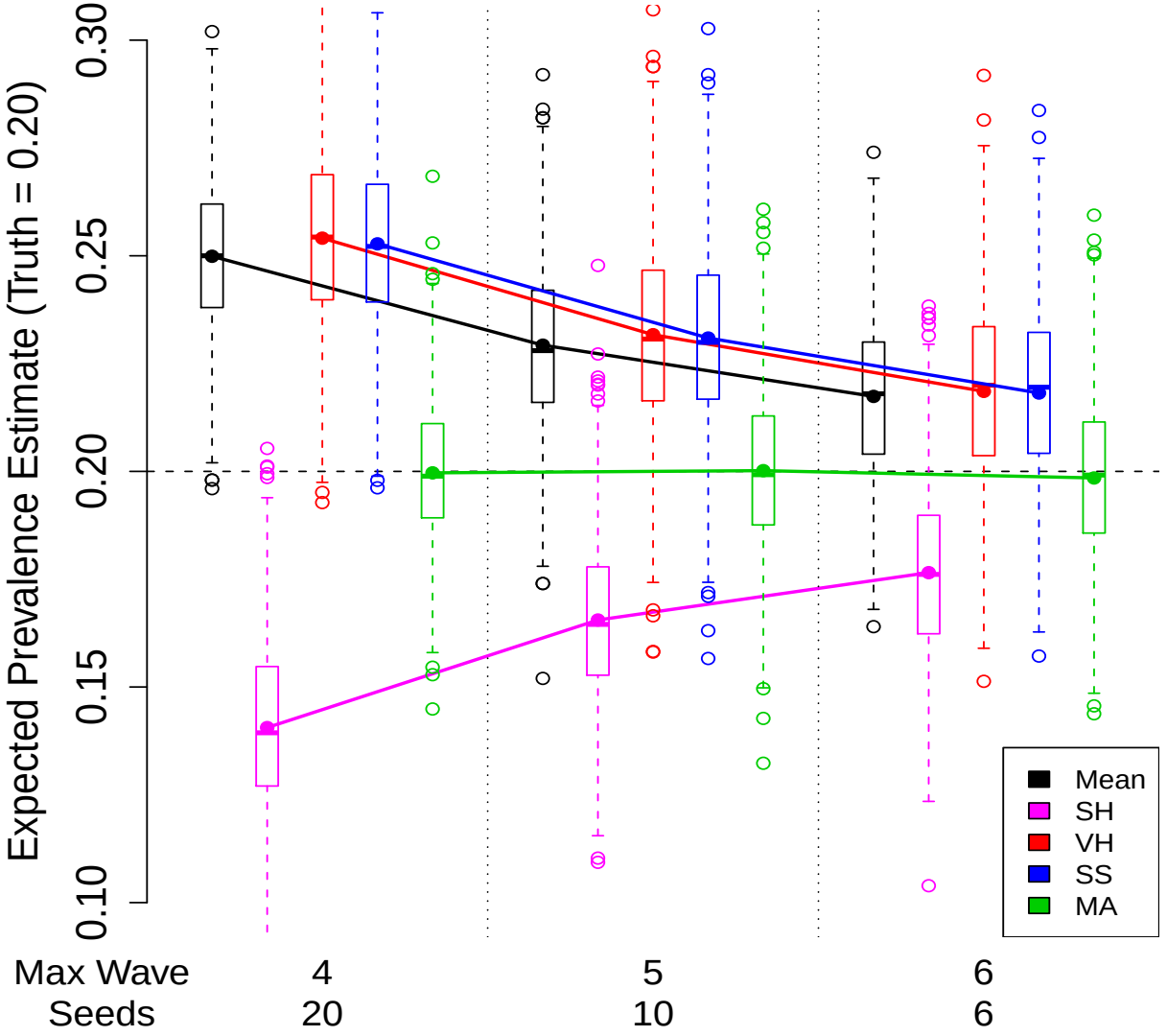
All Infected Seeds, varying number of seeds (waves)



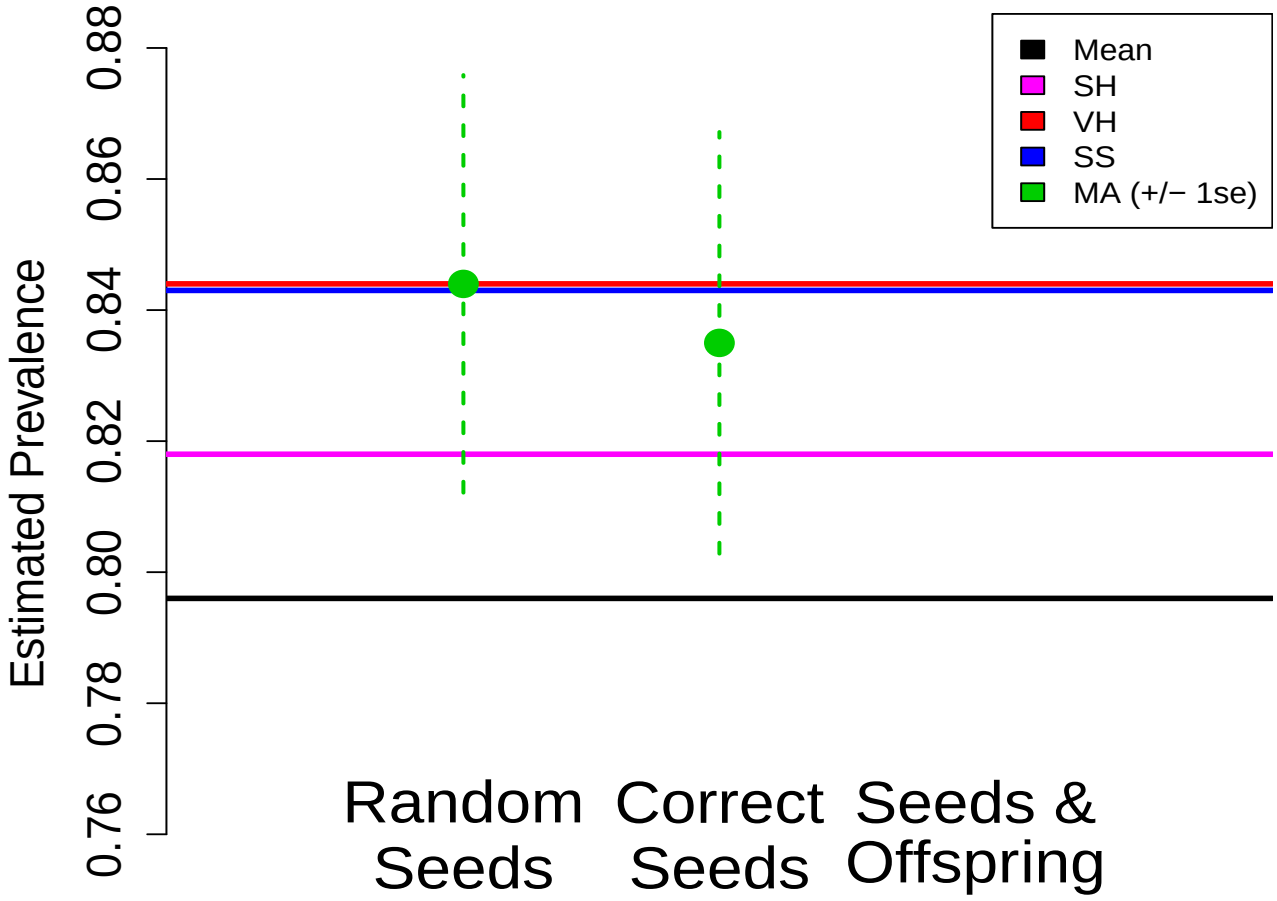
All Infected Seeds, varying number of seeds (waves)



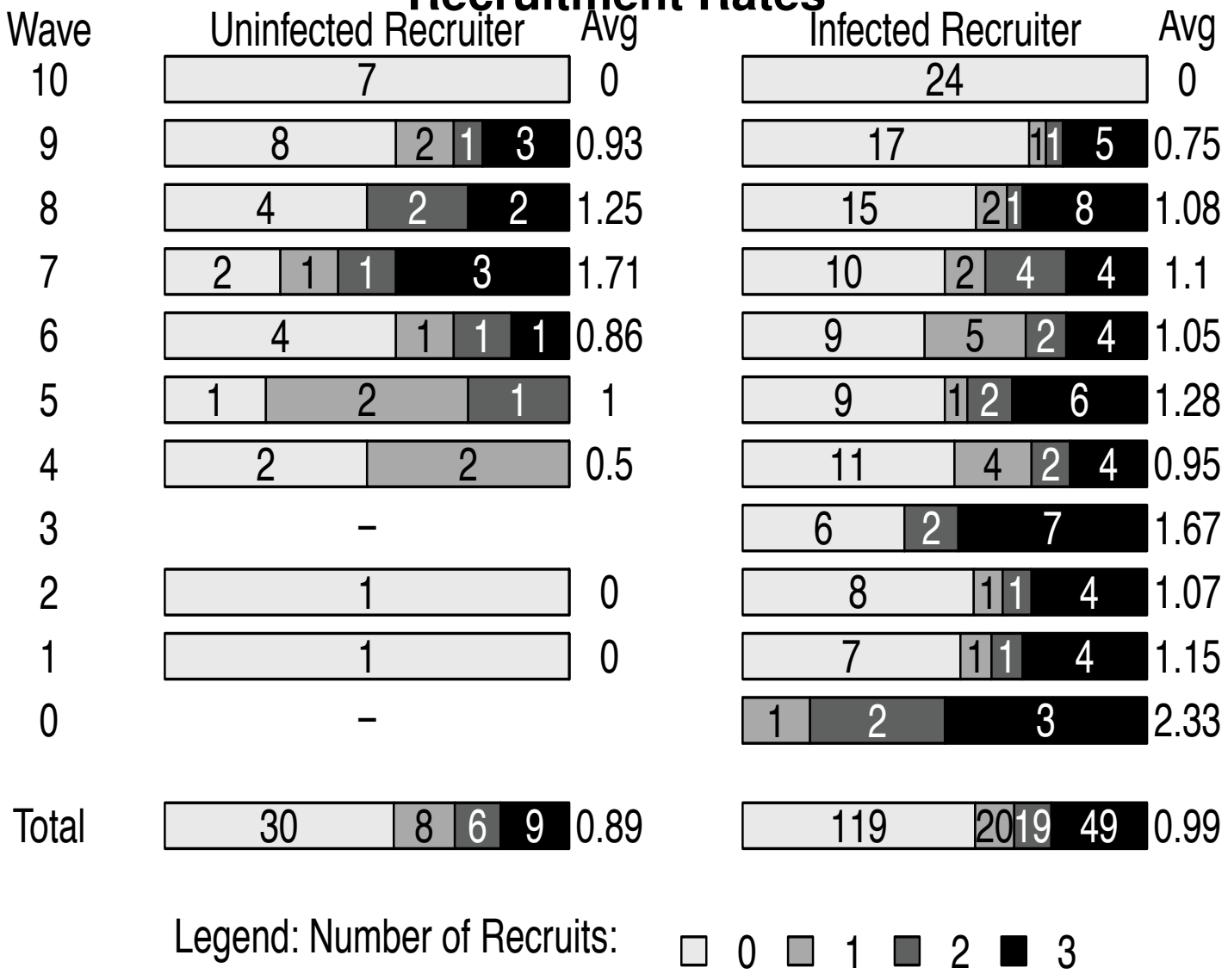
All Infected Seeds, varying number of seeds (waves)



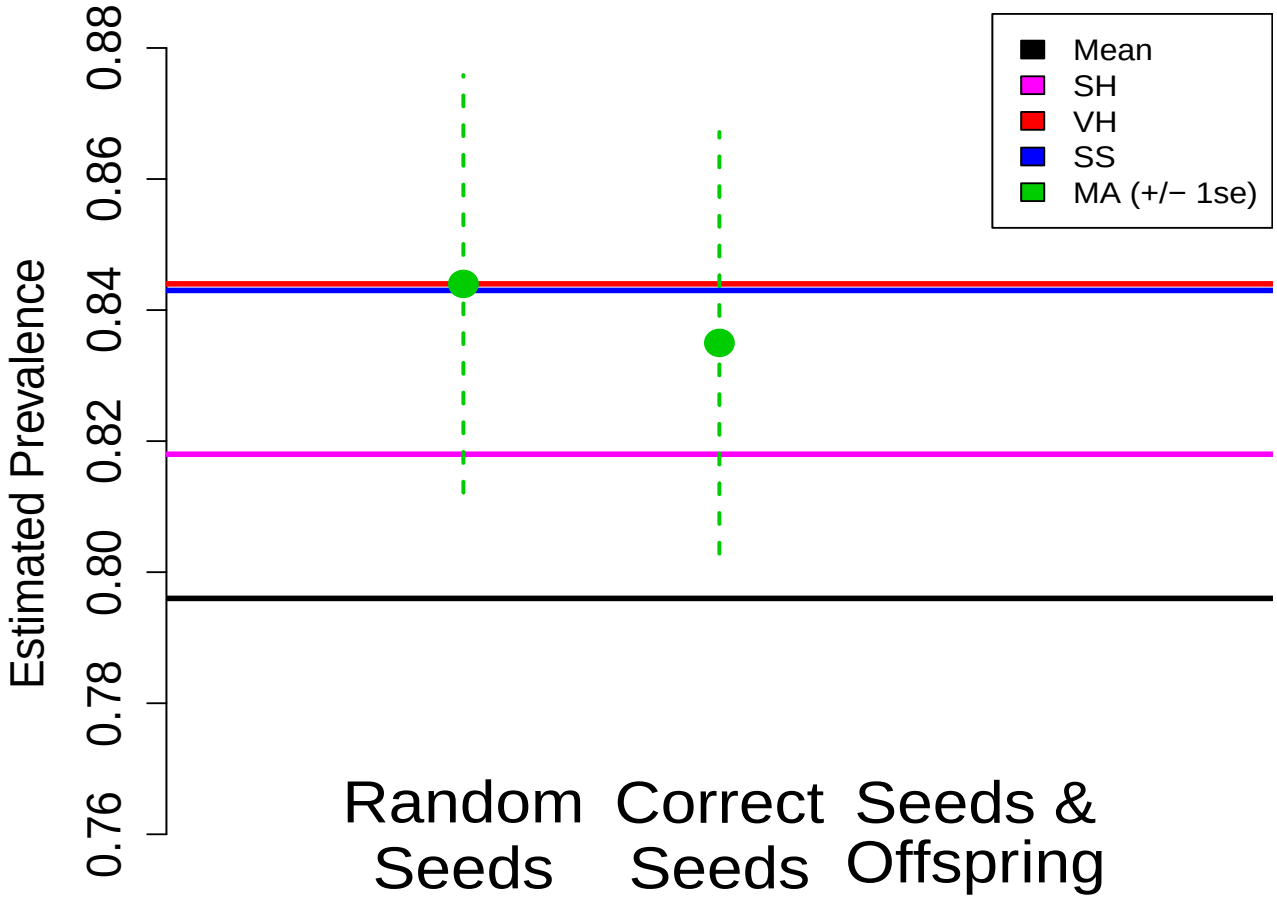
HIV Prevalence among IDU in an Eastern European City



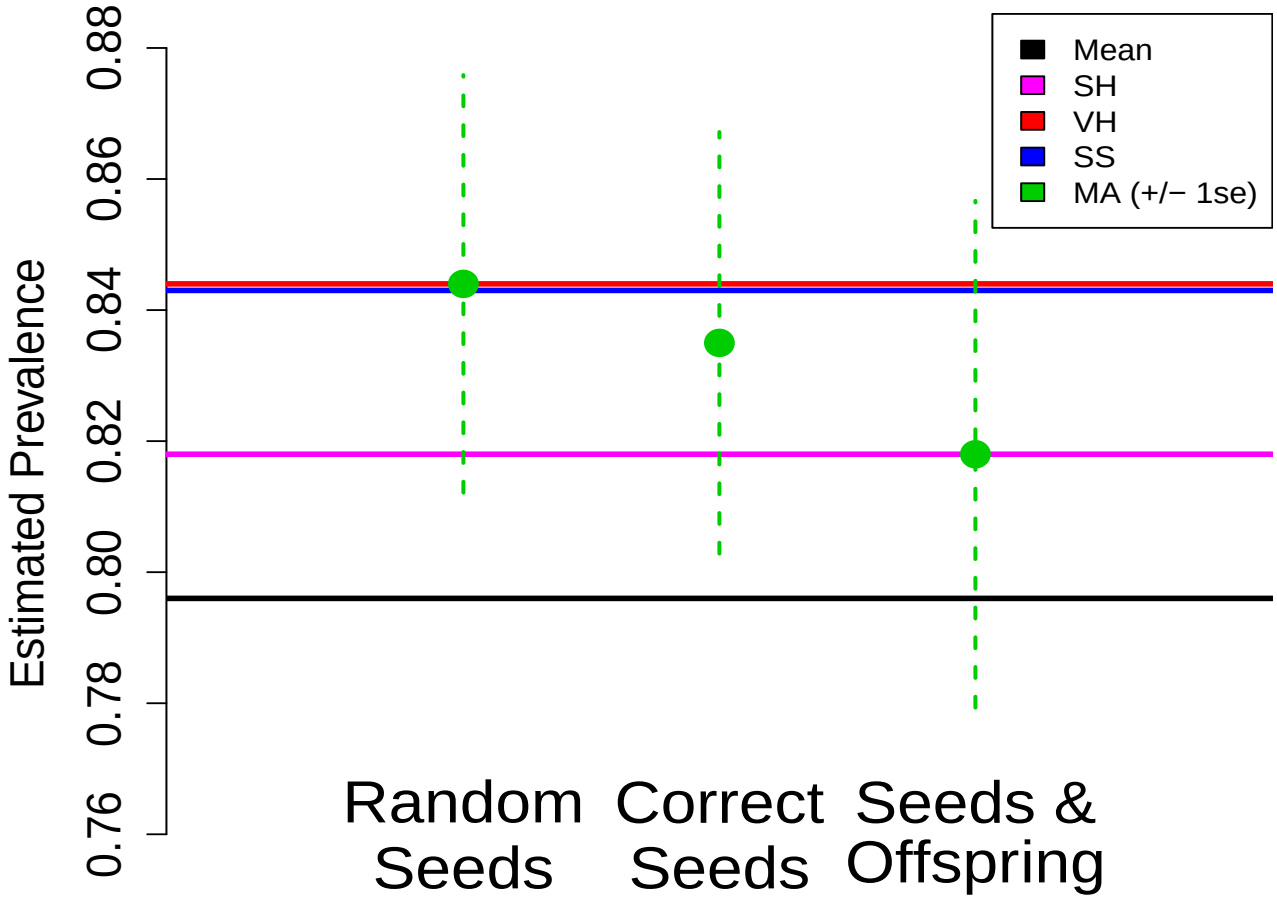
Recruitment Rates



HIV Prevalence among IDU in an Eastern European City



HIV Prevalence among IDU in an Eastern European City



Methodological

- Network-Model estimator corrects for differential activity by infection status, unlike sample mean.
- Network-Model estimator uses appropriate sample weights for simulated high sample fraction, unlike sample mean or Volz-Heckathorn estimator.
- Network-Model estimator corrects for seed bias, unlike any existing method.

Outline of Presentation

1. What is a Social Network?
2. Why do we want to analyze a social network?
3. What can we know about a social network?
4. How do we analyze a social network?
5. Descriptive Analysis
6. Design-Based Inference
7. Model-based (likelihood) Inference
8. Examples
9. Discussion

Discussion

- Examples
  - School Friendships: Describe process generating relations. Missing data
  - Injecting Drug User: Describe nodes in actual population. Network for sampling
- Network data have many types of complexity (nested nodes/edges, time, attributes, boundaries, sampling
- No single dominant approach
- Can only correct for what we can measure - data and scientific question related
- Room for future research: deeper, also broader.

References

- **Missing Data and Sampling**
  - Little, R. J.A. and D. B. Rubin, Second Edition (2002). *Statistical Analysis with Missing Data*, John Wiley and Sons, Hoboken, NJ.
  - Thompson, S.K., and G.A. Seber (1996). *Adaptive Sampling* John Wiley and Sons, Inc. New York.
- **Modeling Social Network Data with Exponential-Family Random Graph Models**
  - Handcock, M.S., D.R. Hunter, C.T. Butts, S.M. Goodreau, and M. Morris (2003) `statnet`: An R package for the Statistical Modeling of Social Networks. URL: `http://www.csde.washington.edu/statnet`.
  - Holland, P.W., and S. Leinhardt (1981), An exponential family of probability distributions for directed graphs, *Journal of the American Statistical Association*, **76**: 33-50.
  - Snijders, T.A.B., P.E. Pattison, G.L. Robins, and M.S. Handcock (2006). New specifications for exponential random graph models. *Sociological Methodology*, 99-153.
- **Inference with Partially-Observed Network Data**
  - Frank, O. (1971). *The Statistical Analysis of Networks* Chapman and Hall, London.
  - Frank, O., and T.A.B. Snijders (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, **10**: 53-67.
  - Gile, K.J. (2008). Inference from Partially-Observed Network Data. PhD. Dissertation. University of Washington, Seattle.
  - Gile, K. and M.S. Handcock (2006). Model-based Assessment of the Impact of Missing Data on Inference for Networks. Working paper, Center for Statistics and the Social Sciences, University of Washington.
  - Handcock, M.S., and K. Gile (2007). Modeling social networks with sampled data. Technical Report, Department of Statistics, University of Washington.
  - Thompson, S.K. and O. Frank (2000). Model-Based Estimation With Link-Tracing Sampling Designs. *Survey Methodology* , **26**: 87-98.
- **Other**
  - Harris, K. M., F. Florey, J. Tabor, P. S. Bearman, J. Jones, and R. J. Udry (2003). The National Longitudinal Study of Adolescent Health: Research design. Technical Report, Carolina Population Center, University of North Carolina at Chapel Hill.

E-mail: gile@math.umass.edu

Thank you for your attention!